全面解读Stata在统计分析中的行业应用

Stata统计分析与行业应用案例详解

·第2版·

（适用范围为Stata 12.0到14.0）

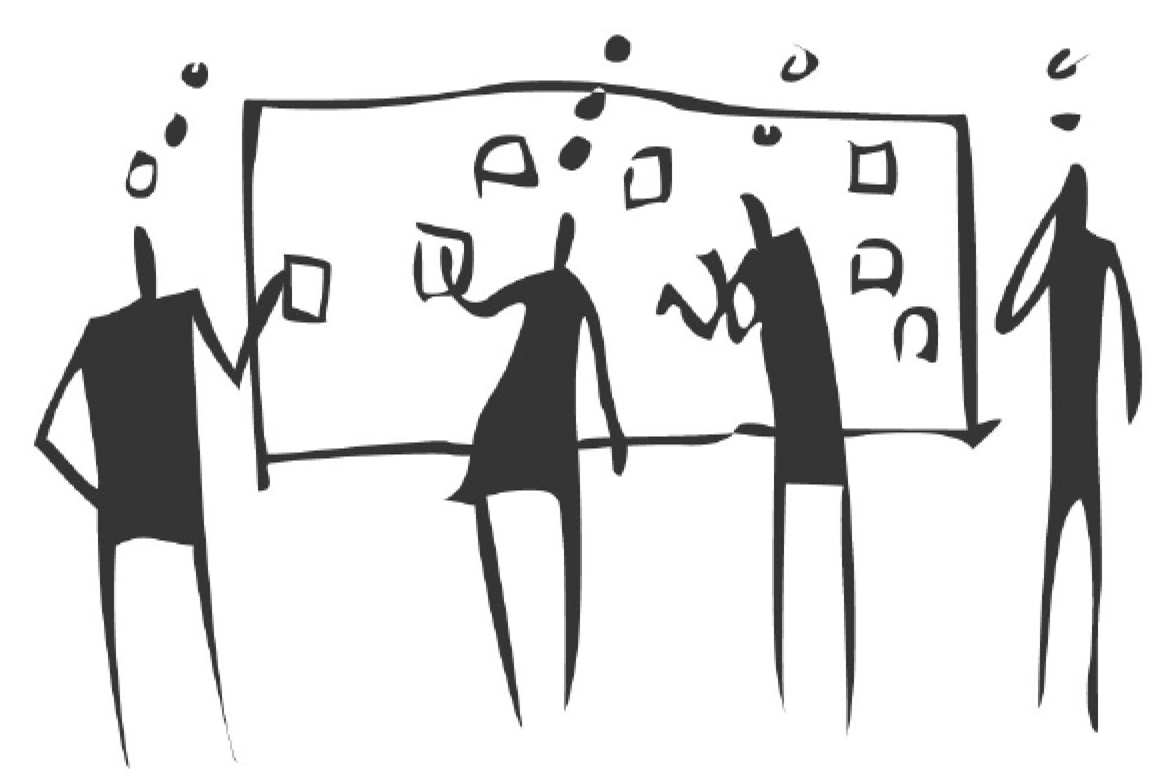


张甜　李爽　编著

57 个基础案例及 **7** 个大型行业应用案例详解Stata统计分析方法、思路和分析流程
61 个上机练习让读者学练结合，快速掌握Stata统计分析方法



清華大学出版社



Stata

统计分析与行业应用案例详解

·第2版·

清華大学出版社
北京

内 容 简 介

Stata是公认的应用最广泛的专业数据分析软件之一，因其功能丰富、效率高、操作简便，深受广大用户，尤其受在校师生的青睐。

本书为《Stata统计分析与行业应用案例详解》的升级版本（Stata 14.0），沿用第一版（Stata 12.0）的写作风格，采用先讲解Stata的各个操作功能再通过综合案例讲述Stata在各个行业中实际应用的思路编写。本书内容共分为两个部分：第1部分是第1~16章，按照统计类型讲述Stata的具体应用；第2部分是第17~23章，分行业讲述了Stata的具体应用。各章均附有与正文部分对应的上机操作练习题，目的是着重培养读者的动手能力，使读者在实际练习的过程中能够快速提高应用水平。

本书面向具备一定统计学基础和计算机操作基础的在校各专业学生，以及企事业单位的相关数据统计分析人员。

图书在版编目（CIP）数据

Stata 统计分析与行业应用案例详解 / 张甜，李爽编著. —2 版. — 北京：清华大学出版社，2017
ISBN 978-7-302-48163-8

I. ①S… II. ①张… ②李… III. ①统计分析－应用软件－案例 IV. ①C819

中国版本图书馆 CIP 数据核字（2017）第 208515 号

责任编辑：夏毓彦
封面设计：王　翔
责任校对：闫秀华
责任印制：
出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦A座　　**邮　　编**：100084
社 总 机：010-62770175　　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn

印 装 者：
经　　销：全国新华书店
开　　本：190mm×260mm　　**印　　张**：36.75　　**字　　数**：941千字
版　　次：2017年9月第1版　　**印　　次**：2017年9月第1次印刷
印　　数：1~3000
定　　价：99.00元

产品编号：068679-01

前 言

Stata 是公认的应用最广泛的专业数据分析软件之一，以功能丰富、效率高、操作简便而著称，主要针对经济、管理、医学、农学、教育、市场研究、社会调查等多个行业和领域。Stata 拥有最具亲和力的窗口，使用者自行建立程序时，软件能提供具有直接命令式的语法，是非常适合进行数据分析的工具软件。本书在第一版 Stata 12.0 的基础上进行了软件版本升级，通过多个实例详细介绍了 Stata 14.0 在现实生活中的应用。

全书共 23 章，分为如下两个部分。

第 1 部分（第 1~16 章）为 Stata 的各个操作功能在具体实例中的应用。

- 第 1 章介绍了 Stata 14.0 基本窗口以及管理变量与数据，包括 Stata 14.0 窗口说明、数据文件的创建与读取、创建和替代变量、分类变量和定序变量的基本操作、数据的基本操作以及定义数据的子集等。
- 第 2 章介绍了 Stata 制图实例，包括直方图、散点图、曲线标绘图、连线标绘图、箱图、饼图、条形图、点图等。
- 第 3 章介绍了 Stata 描述统计实例，包括定距变量的描述性统计分析、正态性检验和数据转换、单个分类变量的汇总、两个分类变量的列联表分析、多表和多维列联表分析等。
- 第 4 章介绍了 Stata 参数检验实例，包括单一样本 T 检验、独立样本 T 检验、配对样本 T 检验、单一样本方差的假设检验、双样本方差的假设检验等。
- 第 5 章介绍了 Stata 非参数检验实例，包括单一样本的正态分布检验、两独立样本检验、两相关样本检验、多独立样本检验、游程检验等。
- 第 6 章介绍了 Stata 方差分析实例，包括单因素方差分析、多因素方差分析、协方差分析、重复测量方差分析等。
- 第 7 章介绍了 Stata 相关分析实例，包括简单相关分析、偏相关分析等。
- 第 8 章介绍了 Stata 主成分分析与因子分析实例。
- 第 9 章介绍了 Stata 聚类分析实例，包括划分聚类分析和层次聚类分析等。
- 第 10 章介绍了 Stata 最小二乘线性回归分析实例，包括简单线性回归和多重线性回归等。
- 第 11 章介绍了 Stata 回归诊断分析实例，包括异方差检验、自相关检验、多重共线性检验等。
- 第 12 章介绍了 Stata 非线性回归分析实例，包括非参数回归分析、转换变量回归分析以及非线性回归分析等。
- 第 13 章介绍了 Stata 的 Logistic 回归分析实例，包括二元 Logistic 回归分析、多元 Logistic 回归分析以及有序 Logistic 回归分析等。
- 第 14 章介绍了 Stata 的因变量受限回归分析实例，包括断尾回归分析和截取回归分析。

- 第 15 章介绍了 Stata 时间序列分析实例，包括时间序列分析的基本操作、单位根检验、协整检验、格兰杰因果关系检验等。
- 第 16 章介绍了 Stata 的面板数据分析实例，包括长面板数据分析和短面板数据分析。

第 2 部分（第 17~23 章）为 Stata 在各个行业中的实际应用。

- 第 17 章介绍了 Stata 在研究城市综合经济实力中的应用。
- 第 18 章介绍了 Stata 在旅游业中的应用。
- 第 19 章介绍了 Stata 在经济增长分析中的应用。
- 第 20 章介绍了 Stata 在原油与黄金价格联动关系研究中的应用。
- 第 21 章介绍了中国上市银行的 ROE 与股权集中度之间关系研究中的应用。
- 第 22 章介绍了 Stata 在农业中的应用。
- 第 23 章介绍了 Stata 软件在保险业中的应用。

本书实例经典，内容丰富，有很强的针对性。书中各章不仅详细介绍了实例的具体操作步骤，还配有一定数量的练习题，以供读者学习使用。读者只需按照书中介绍的步骤一步步地实际操作，就能完全掌握本书的内容。

为了帮助读者更加直观地学习本书，我们将书中实例和练习题所涉及的全部操作文件都收录到本书的下载资源中，即“sample”文件夹和“video”文件夹。前者包含书中涉及的所有 Stata 源文件，后者收录了书中所有实例和练习题的操作录像文件。下载资源地址为：http://pan.baidu.com/s/1cejAHK（注意区分字母的大小写及数字和字母，若下载有疑问，可发邮件至 booksaga@163.com）。

本书既可作为数据统计分析的培训教材，也可作为数据统计分析人员的参考书。

本书由张甜、李爽编写，此外，参与图书编写和视频制作的还有吕平、王坚宁、高克臻、张云霞、许小荣、王冬、王龙、张银芳、周新国、张凤琴、陈作聪、聂阳、沈毅、张华杰、彭一明、张秀梅、张玉兰、田伟、肖岳平、蔡娜、苏静、周艳丽和王文婷等，在这里对他们表示感谢。

作者力图使本书的知识性和实用性相得益彰，但由于水平有限，书中纰漏之处在所难免，欢迎广大读者、同仁批评斧正。

编　者

2017 年 3 月

目　录

第 1 章　Stata 14.0 的基本窗口及管理变量与数据

Stata 是一种功能全面的统计软件包，是目前欧美最为流行的计量软件之一。它具有容易操作、运行速度快、功能强大的特点。Stata 不仅包括一整套预先编排好的分析与数据功能，同时还允许软件使用者根据自己的需要来创建程序，从而添加更多的功能。该软件自从被引入我国后，迅速得到了广大学者的认可与厚爱，适用范围越来越广泛。Stata 14.0 是目前 Stata 的最新版本。本章将初步介绍 Stata 14.0 的基本窗口、变量管理与数据管理。

1.1　Stata 14.0窗口说明

在正确安装好 Stata 14.0 以后，单击 Stata 主程序的图标文件，即可打开 Stata 的主界面，如图 1.1 所示。

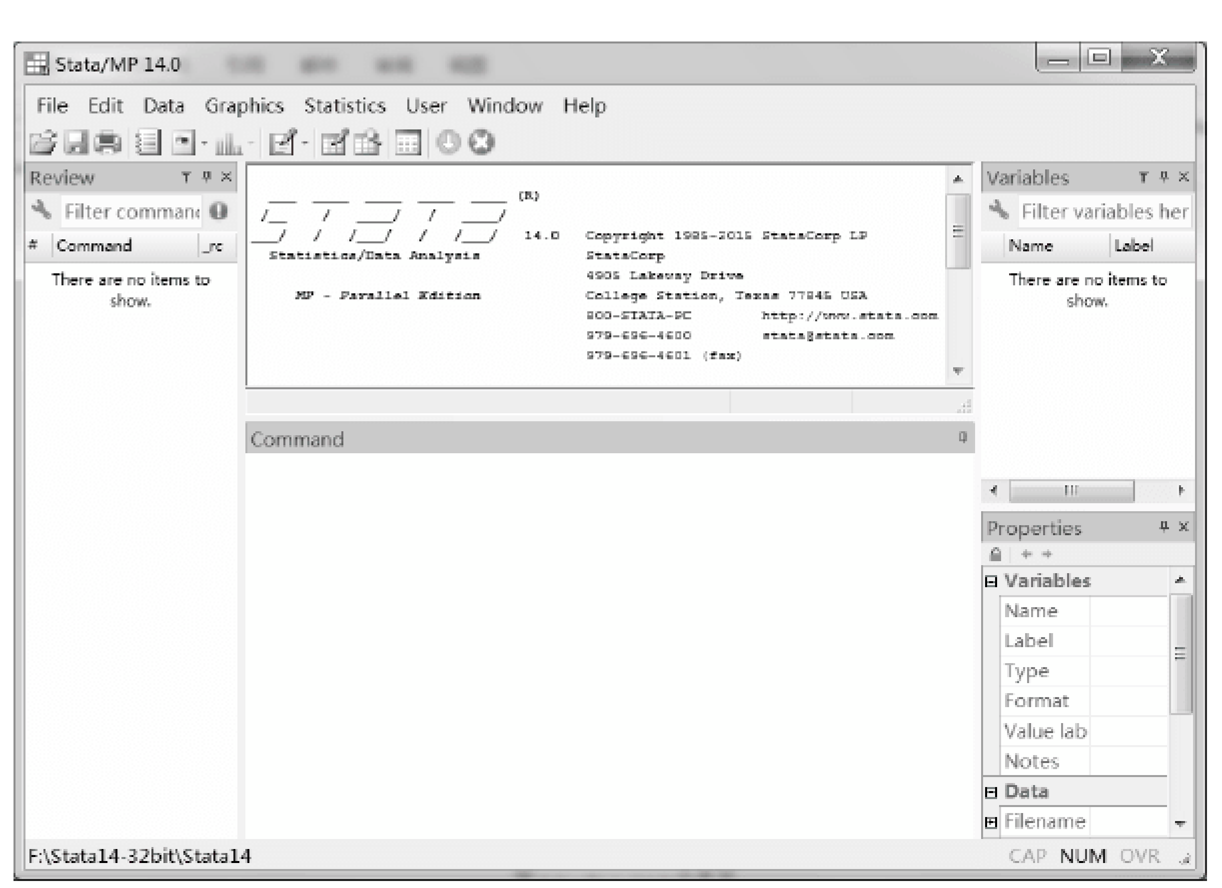


图 1.1　Stata 14.0 主界面

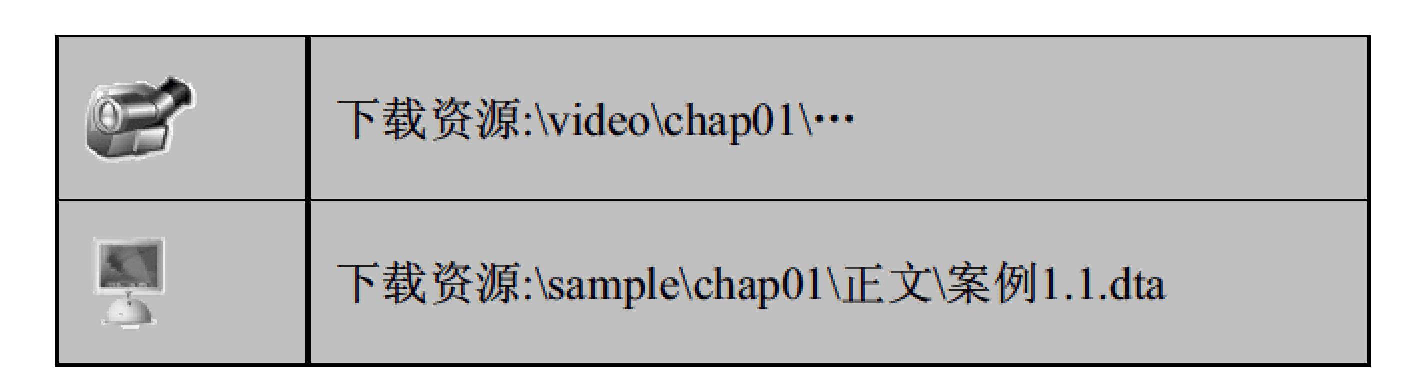
与大部分的程序窗口类似，Stata 14.0 也有自己的菜单栏、工具栏，但其特色在于主界面中的 5 个区域：Review、Variables、Command、Results、Properties。

- Review（历史窗口）显示的是自本次启动 Stata 14.0 以来执行过的所有命令。
- Variables（变量窗口）显示的是当前 Stata 数据文件中的所有变量。
- Command（命令窗口）是最重要的窗口，在本窗口内可输入准备执行的命令。
- Results（结果窗口）显示的是每次执行 Stata 命令后的执行结果，无论成功还是失败。
- Properties（性质窗口）显示的是当前数据文件中制定变量以及数据的性质。

各个窗口的大小都可以调节，读者可以用鼠标进行伸缩操作，使其符合自己的风格。

1.2　Stata 14.0数据文件的创建与读取

1.2.1　Stata 14.0 数据文件的创建



	下载资源:\video\chap01\···
	下载资源:\sample\chap01\正文\案例1.1.dta

【例 1.1】表 1.1 记录的是我国 2000—2009 年上市公司数量的数据。试创建 Stata 格式的数据文件并保存。

表 1.1　我国 2000—2009 年的上市公司数量

年份	上交所	深交所
2000	572	516
2001	646	514
2002	715	509
2003	780	507
2004	837	540
2005	834	547
2006	842	592
2007	860	690
2008	864	761
2009	870	848

操作过程如下：

01 进入 Stata 14.0，打开主程序，弹出如图 1.2 所示的主界面。

02 选择“Data”|“Data Editor”|“Data Editor(Edit)”命令，弹出如图 1.3 所示的“Data Editor(Edit)”对话框。

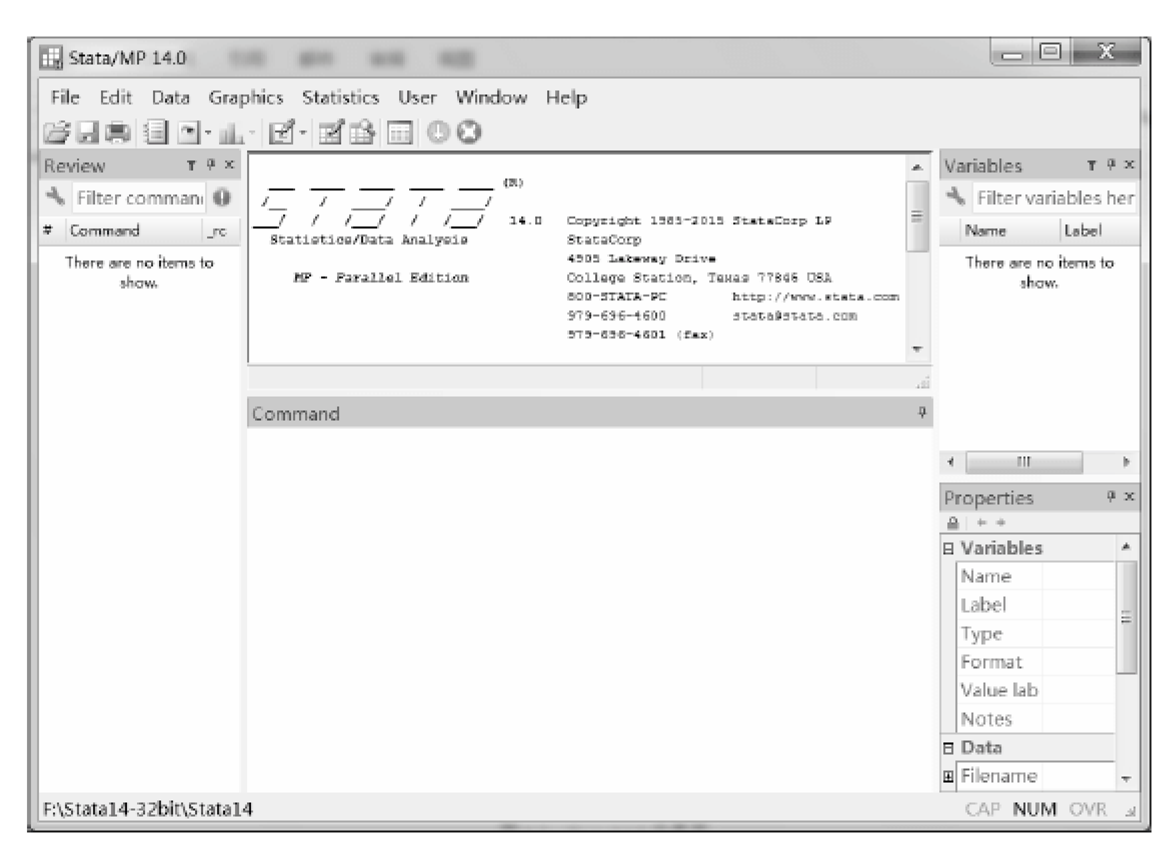


图 1.2　主界面

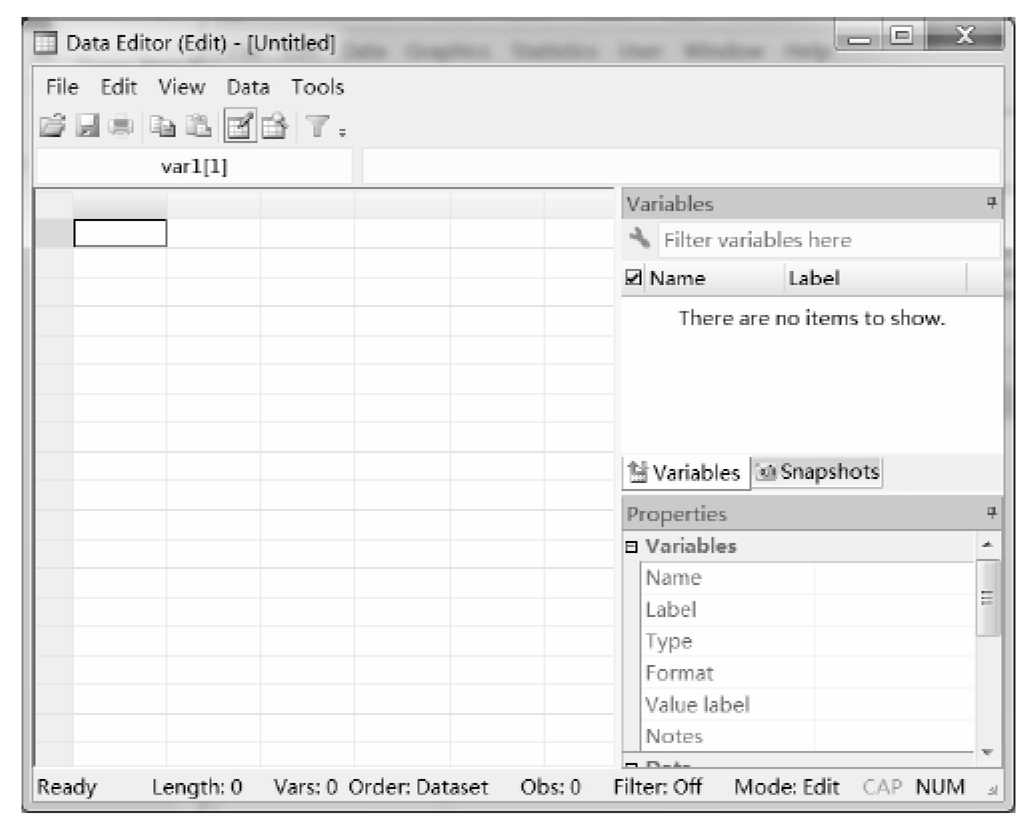


图 1.3　“Data Editor”对话框

03 在“Data Editor(Edit)”对话框左上角的单元格中输入我们的第 1 个数据“2000”，系统即自动创建“var1”变量，如图 1.4 所示。

04 单击右下方“Properties”（性质窗口）中的“Variables”，“Variables”中的变量特征（包括名称、类型、长度等）即可进入可编辑状态，如图 1.5 所示。

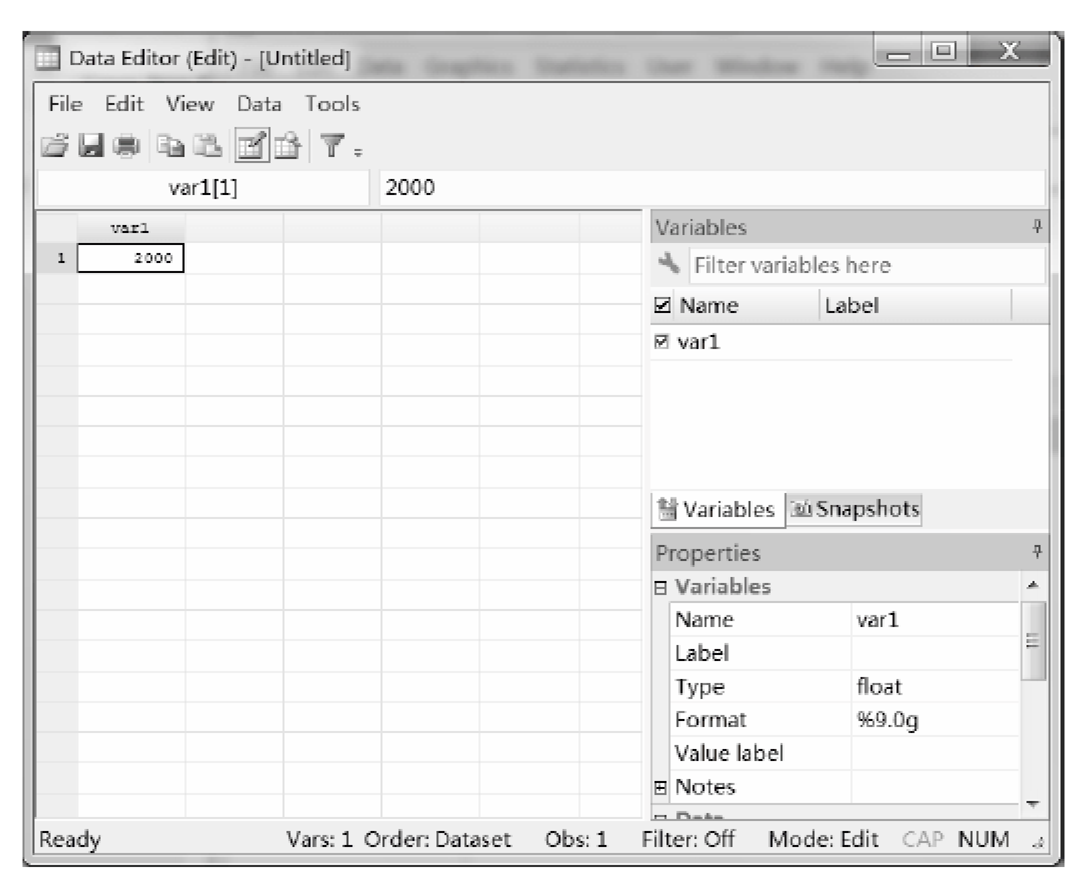


图 1.4　“Data Editor”对话框

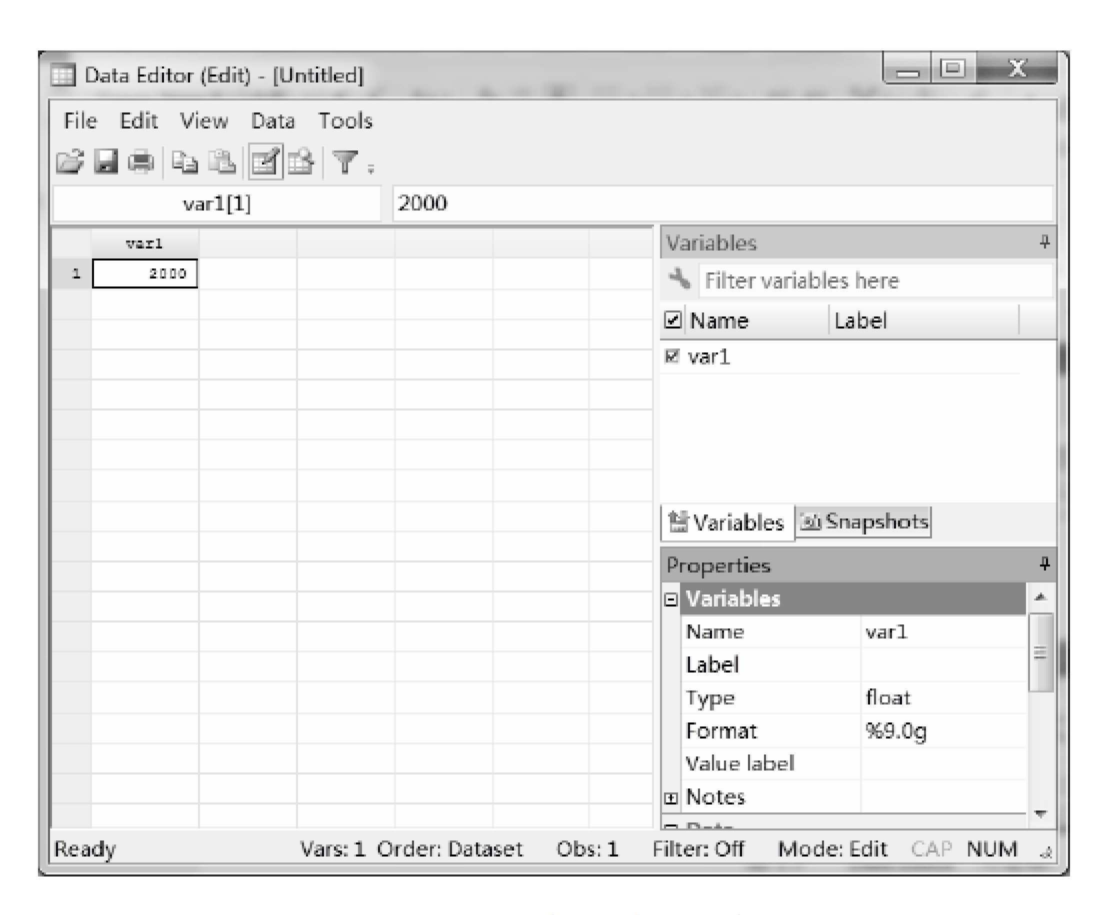
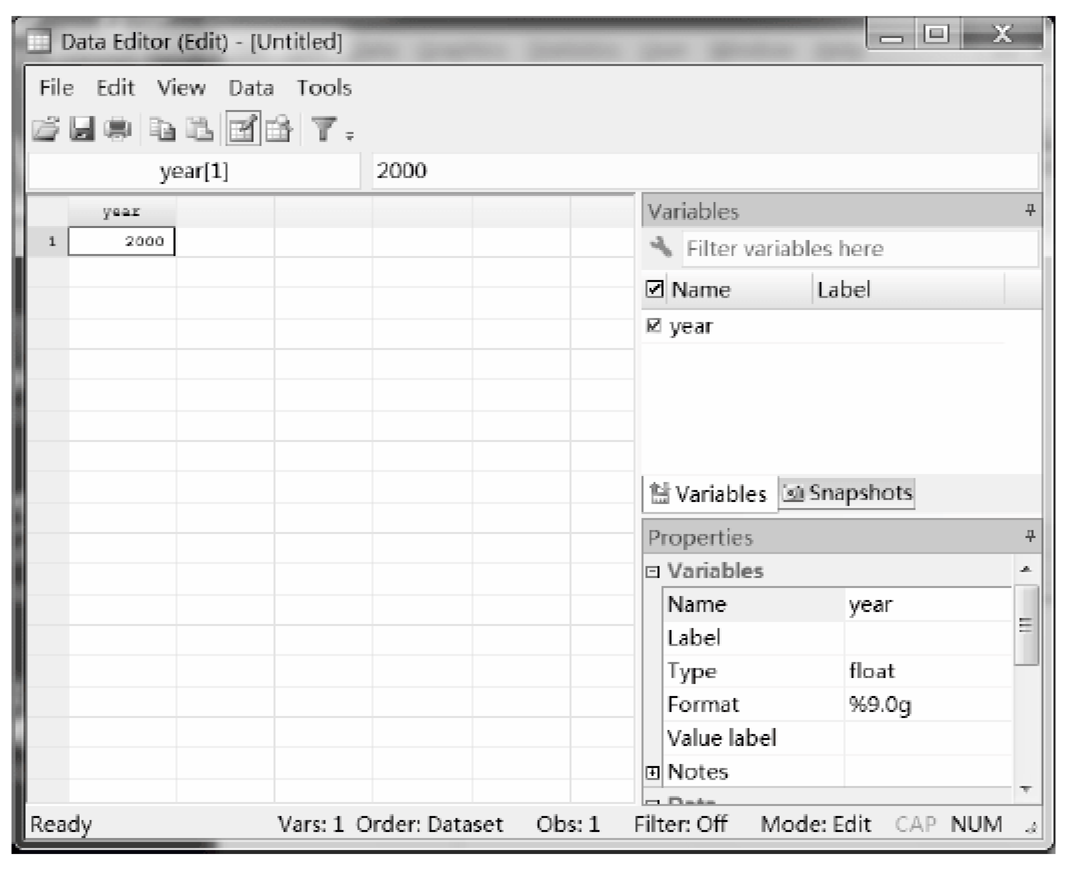


图 1.5　编辑变量特征

05 我们对变量名称进行必要的修改，因为第 1 个变量是年份，所以把“var1”修改为“year”，其他采取系统默认设置，修改完成后在左侧数据输入区域单击，即可弹出如图 1.6 所示的对话框。

06 逐一数据录入，其他两个变量参照年份进行设置，并分别将其定义为“shangjiao”和“shenjiao”，数据录入完毕后如图 1.7 所示。



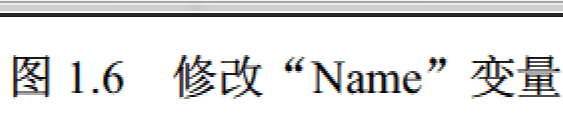


图 1.6　修改“Name”变量

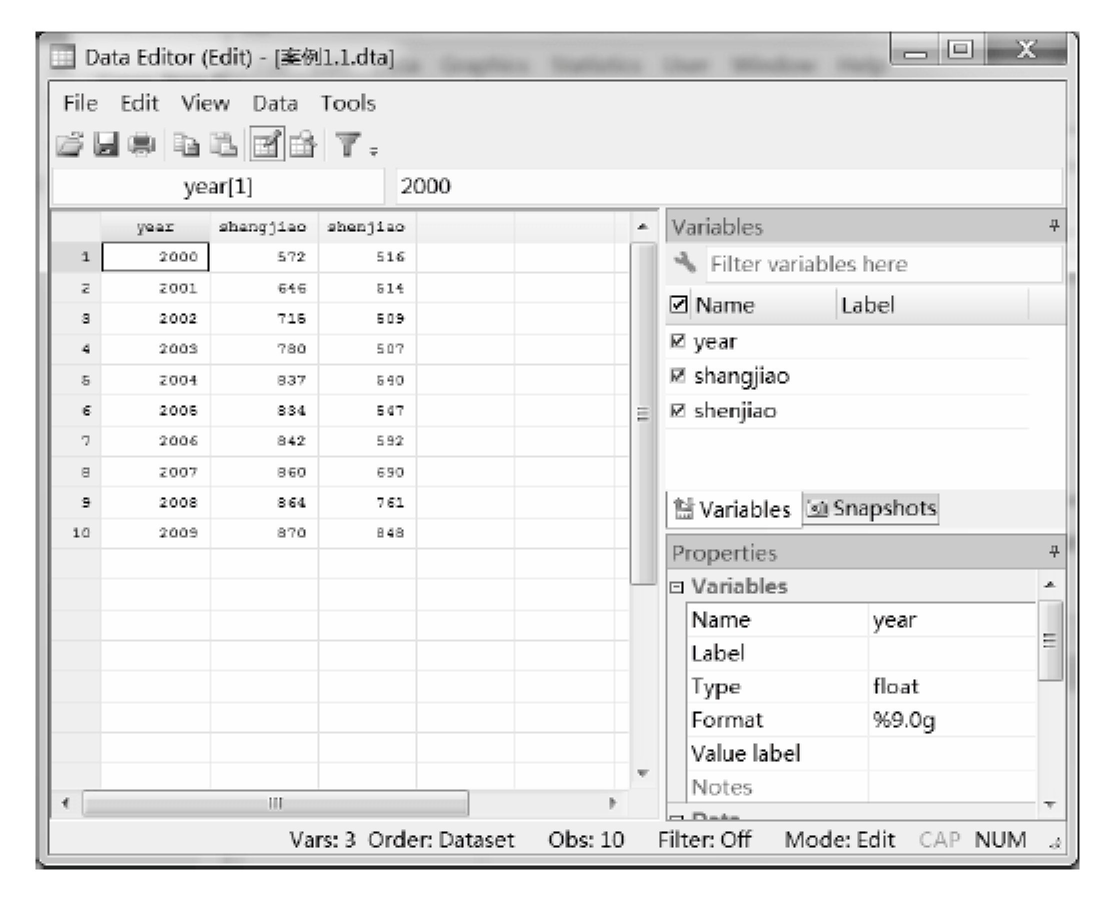


图 1.7　录入数据

07 关闭“Data Editor(Edit)”对话框，在主界面的工具栏里面单击 按钮进行数据保存。

1.2.2　Stata 14.0 数据文件的读取

读取以前创建的 Stata 格式的数据文件比较简单，有 3 种方式：

- 直接双击该文件，即可打开数据。
- 在主界面的菜单栏里面选择“File”|“Open”命令，找到文件后打开即可。
- 在主界面的“Command”（命令窗口）中，输入命令：use filename（文件的名称）。

1.3 创建和替代变量

1.3.1 创建和替代变量概述

前面已经介绍了创建、修改数据文件和变量的通用方式，但在有些情况下，我们需要利用现有的变量生成一个新的变量，那么如何快捷方便地实现这种操作呢？Stata 14.0 提供了 generate 以及 replace 命令以供我们选择使用，其中 generate 命令是利用现有变量生成一个新的变量，并保留原来的变量不变；而 replace 命令则是利用现有变量生成一个新的变量替换原来的变量。下面我们就用实例的方式来讲解一下这两个重要命令的应用。

1.3.2 相关数据来源

	下载资源:\video\chap01\…
	下载资源:\sample\chap01\正文\案例1.2.dta

【例 1.2】我国 2009 年各地区的就业人口以及工资总额数据如表 1.2 所示。请使用 Stata 命令进行操作：（1）试生成新的变量来描述各地区的平均工资情况；（2）试生成平均工资变量来替代原有的工资总额变量；（3）对生成的平均工资变量数据均做除以 10 的处理；（4）对就业人口变量进行对数平滑处理，从而产生新的变量。

表 1.2　我国 2009 年各地区的就业人口及工资总额

地区	就业人口/人	工资总额/千元
北京	6 193 478	354 562 114
天津	2 016 501	88 650 773
河北	5 030 626	139 819 814
山西	3 857 975	107 304 259
内蒙古	2 458 276	76 181 130
…	…	…
青海	506 254	16 361 377
宁夏	581 039	19 536 870
新疆	2 494 187	71 506 764

1.3.3 Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有 3 个变量，分别是地区、就业人口、工资总额。我们把地区变量设定为 region，把就业人口变量设定为 people，把

工资总额变量设定为 sumwage，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作我们在 1.2 节中已有详细讲述。录入完成后，数据如图 1.8 所示。

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出如图 1.9 所示的主界面。

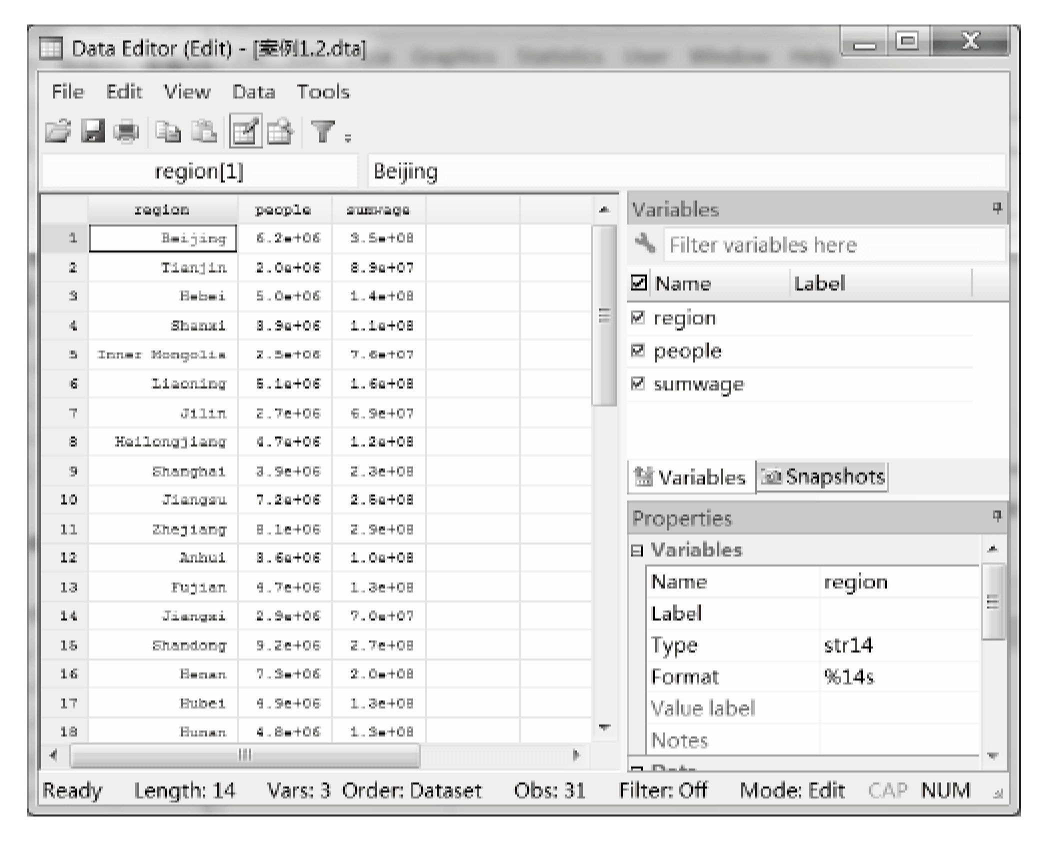


图 1.8　案例 1.2 数据

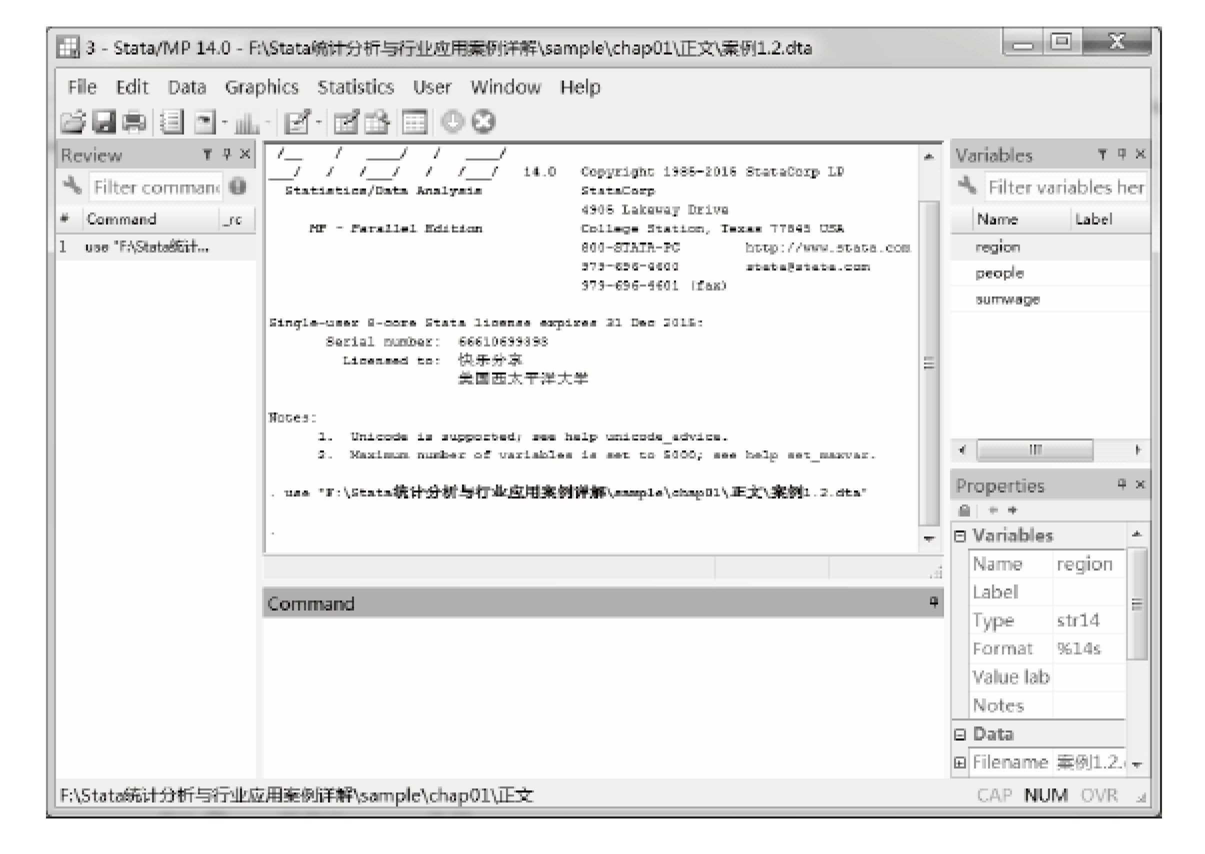


图 1.9　主界面

02 在主界面的“Command”文本框中输入如下操作命令并按键盘上的回车键进行确认。

- generate avwage= sumwage/ people：本命令的含义是生成新的变量来描述各地区的平均工资情况。
- replace sumwage= sumwage/ people：本命令的含义是生成平均工资变量来替代原有的工资总额变量。
- replace sumwage= sumwage/ 10：本命令的含义是对生成的平均工资变量数据均做除以 10 的处理。
- gen lpeople=ln(people)：本命令的含义是对就业人口变量进行对数平滑处理，从而产生新的变量。

03 设置完毕后，按键盘上的回车键，等待输出结果。

1.3.4　结果分析

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 1.10 所示的 avwage 数据。

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 1.11 所示的 sumwage 数据，等于总工资除以总职工数。

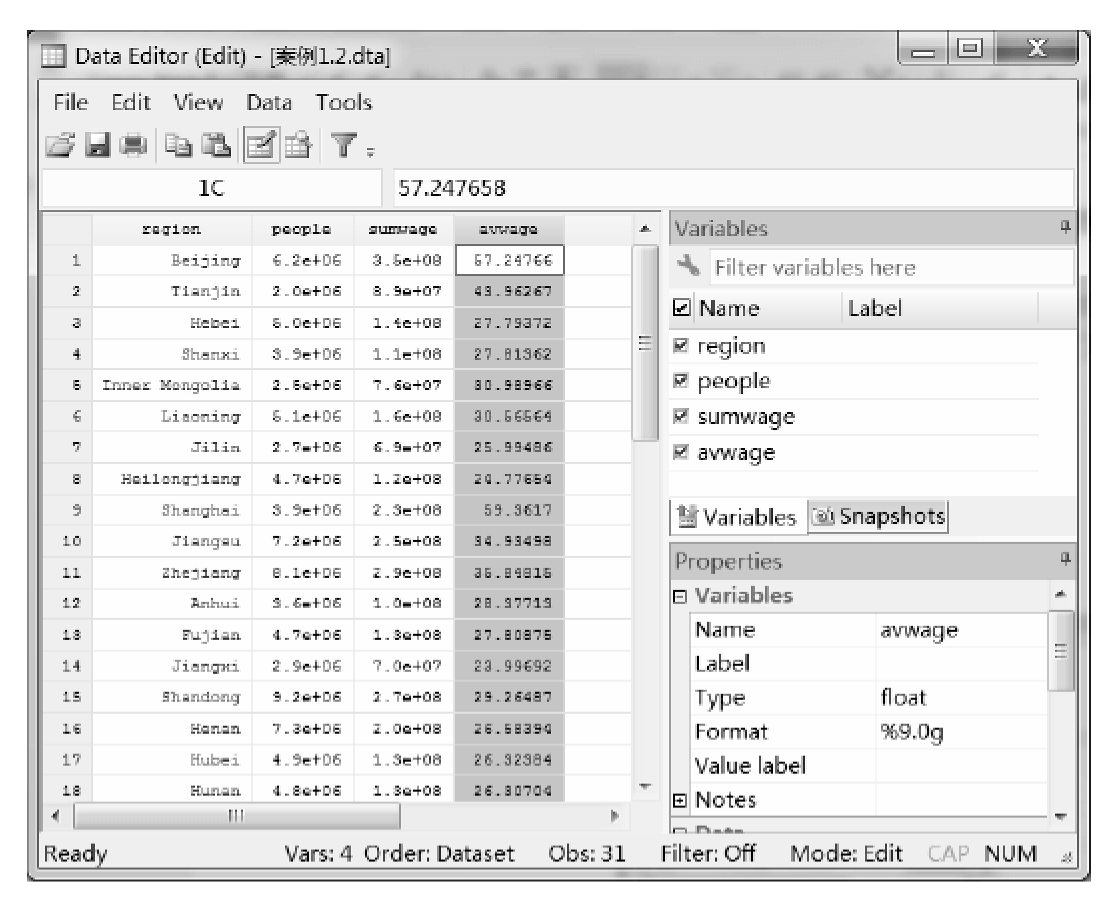


图 1.10　“avwage”数据

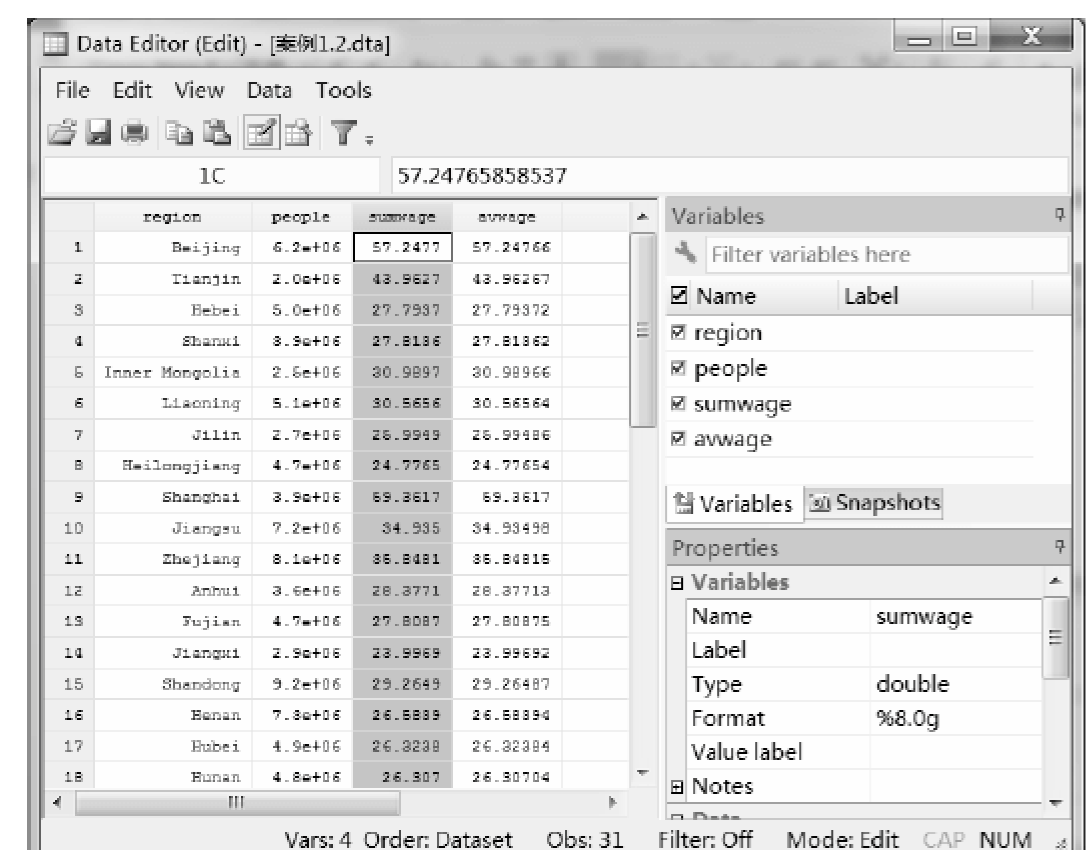


图 1.11　平均工资

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 1.12 所示的 sumwage 数据，即前面生成的平均工资数据除以 10。

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 1.13 所示的 lpeople 数据。它是针对 people 数据取的对数值。

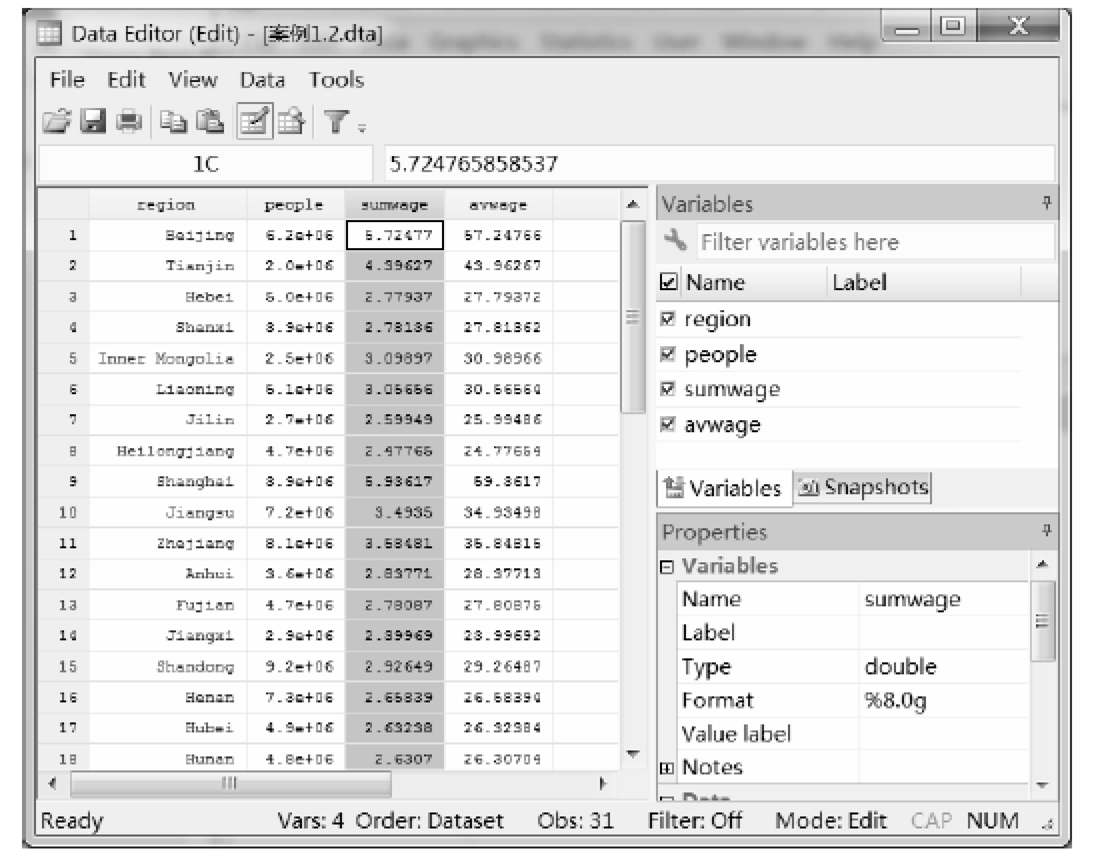


图 1.12　平均工资除以 10

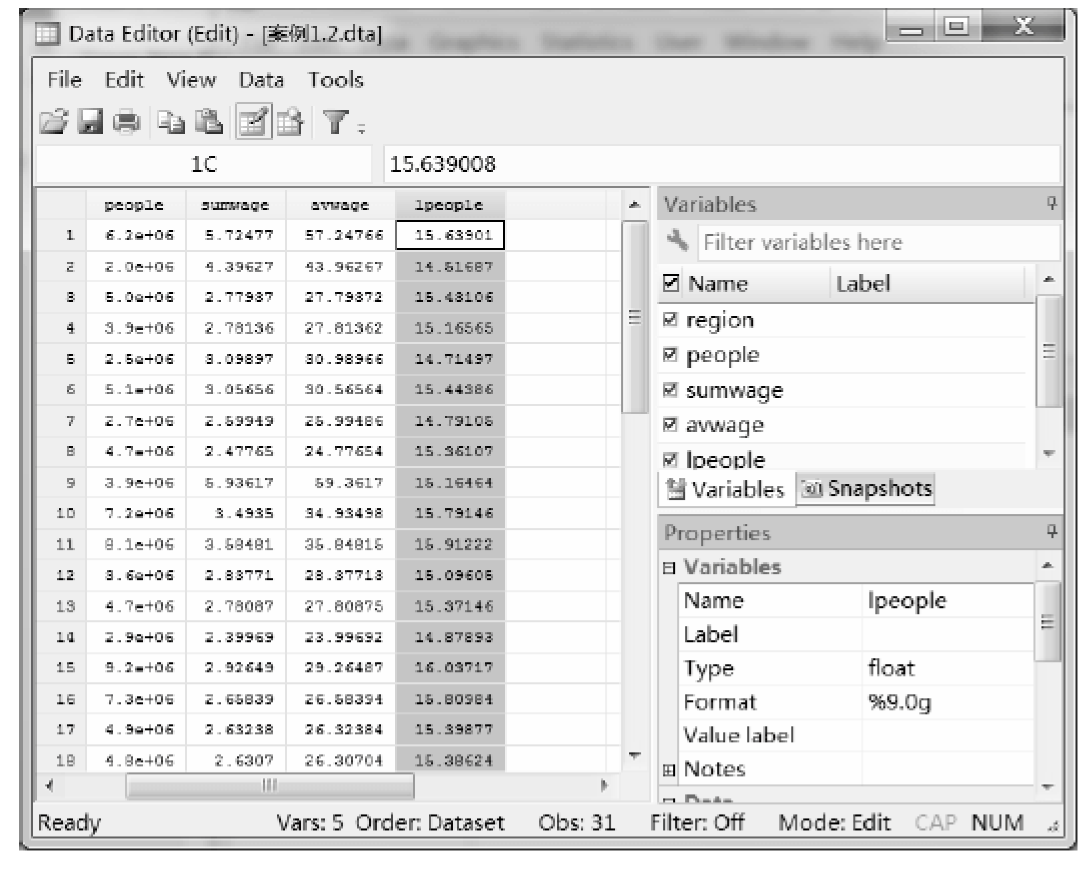


图 1.13　对就业人口进行对数平滑处理

1.3.5　案例延伸

在上面的案例中，我们用到了代数运算符“/”。在 Stata 14.0 中，我们可以使用的代数运算符如表 1.3 所示。

表 1.3　代数运算符

代数运算符	含义	代数运算符	含义	代数运算符	含义	代数运算符	含义	代数运算符	含义
+	加	−	减	*	乘	/	除	^	乘方

在上面的案例中，我们也用到了自然对数函数 ln(变量)。在 Stata 14.0 中，我们经常使用的函数如表 1.4 所示。

表 1.4　函数

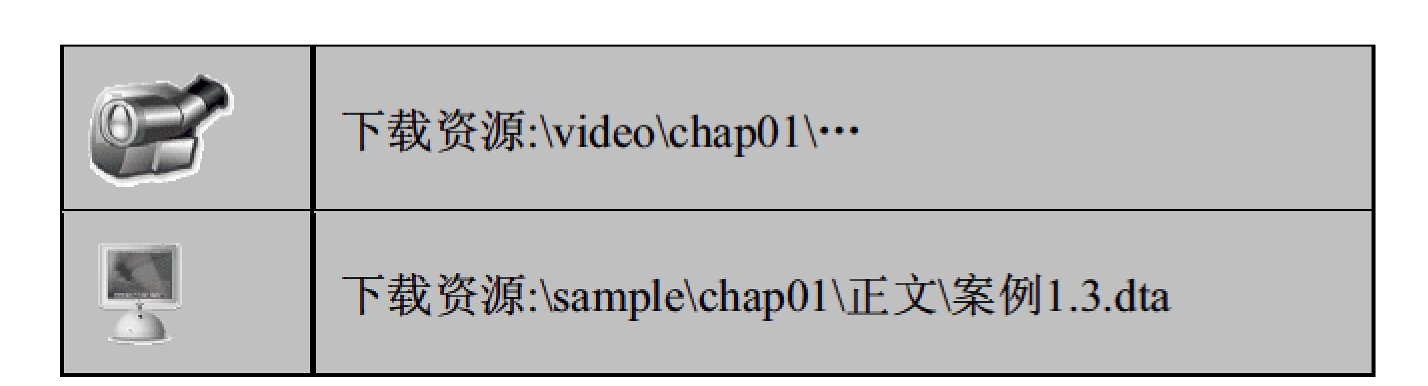
函数命令	表示含义	函数命令	表示含义	函数命令	表示含义
abs(x)	x的绝对值	sqrt	平方根函数	exp(x)	指数函数
sin	正弦函数	cos(x)	余弦函数	tan(x)	正切函数
asin(x)	反正弦函数	acos(x)	反余弦函数	atan(x)	反正切函数
trunk(x)	x的整数部分	logit(x)	x的对数比率	total(x)	x的移动合计
mod(x,y)	x/y的余数	sign(x)	符号函数	round(x)	x的四舍五入整数
atanh(x)	双曲反正切函数	floor(x)	小于等于x的最大整数	ceil(x)	小于等于x的最小整数

1.4　分类变量和定序变量的基本操作

1.4.1　分类变量和定序变量概述

在很多情况下，我们会用到分类变量（虚拟变量）的概念，分类变量（虚拟变量）的用途是通过定义值的方式将观测样本进行分类。例如，根据数据某一变量特征的不同把观测样本分为 3 类，就需要建立 3 个分类变量 A、B、C，如果观测样本属于 A 类，其对应的分类变量 A 的值就为 1，对应的分类变量 B 和 C 的值就为 0。定序变量的用途是根据数据的数值大小将数据分到几个确定的区间，其在广义上也是一种分类。下面我们就用实例的方式来讲解一下分类变量和定序变量的基本操作。

1.4.2　相关数据来源



【例 1.3】某国际知名足球裁判自执法以来在各地区的执赛信息如表 1.5 所示。试使用 Stata 14.0 对数据进行以下操作：（1）试生成新的分类变量来描述比赛级别；（2）试生成新的定序变量对场数进行定序，分到 3 个标志区间。

表 1.5　某国际知名足球裁判执赛情况

地点	场数	比赛级别
江苏	20	省级
浙江	14	省级
安徽	4	省级
福建	3	省级
江西	5	省级
山东	21	省级
美国	10	国家级
日本	19	国家级
英国	32	国家级
挪威	3	国家级

1.4.3 Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有 3 个变量，分别是地点、场数以及比赛级别。我们把地点变量设定为 place，把场数变量设定为 number，把比赛级别变量设定为 type，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作我们在 1.2 节中已有详细讲述。录入完成后数据如图 1.14 所示。

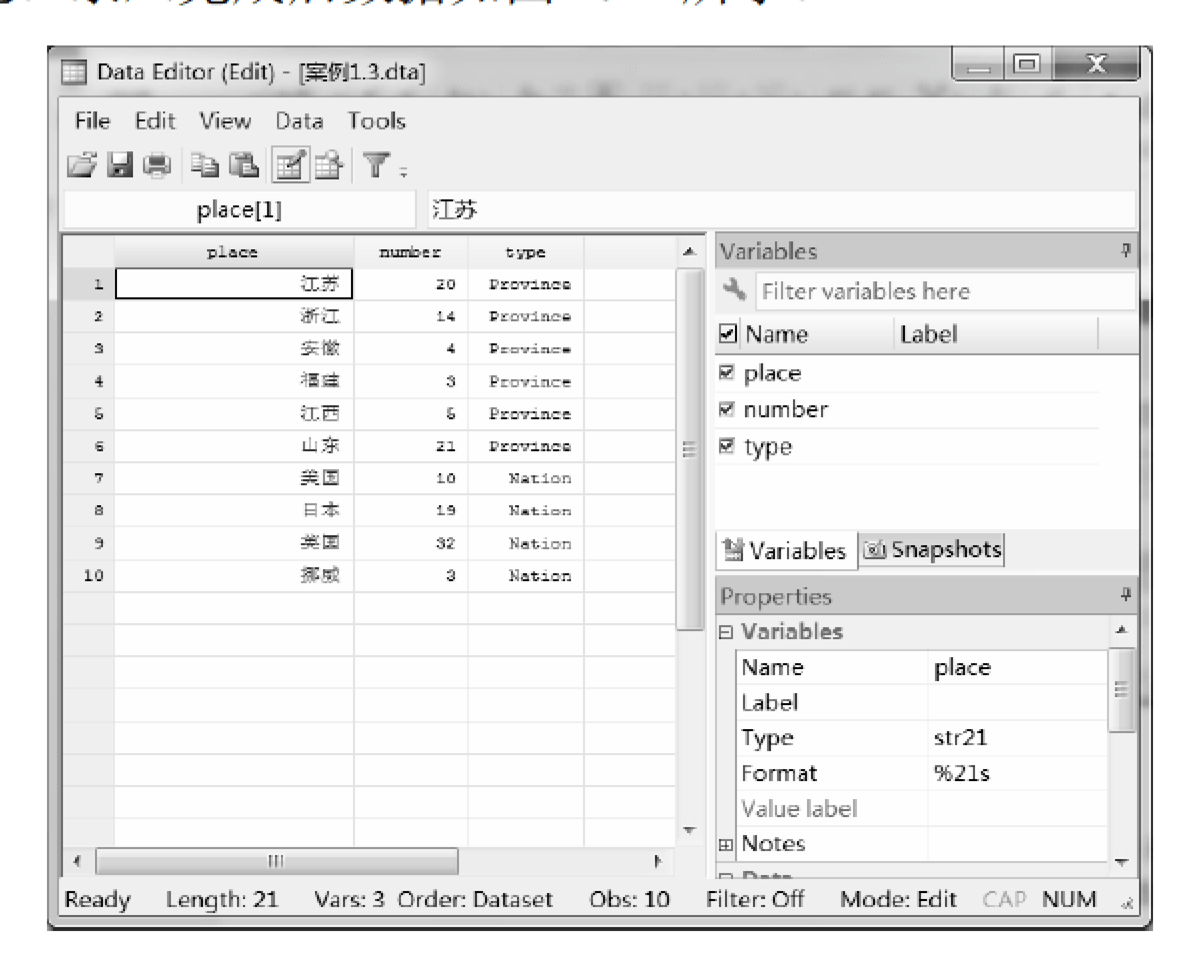


图 1.14 案例 1.3 数据

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出如图 1.15 所示的主界面。

02 在主界面的“Command”文本框中输入操作命令并按键盘上的回车键进行确认。

- tabulate type,generate(type)：本命令的含义是生成新的分类变量来描述比赛级别。
- generate number1=autocode(number,3,1,25)：本命令的含义是生成新的定序变量对场数进行定序，分到 3 个标志区间。

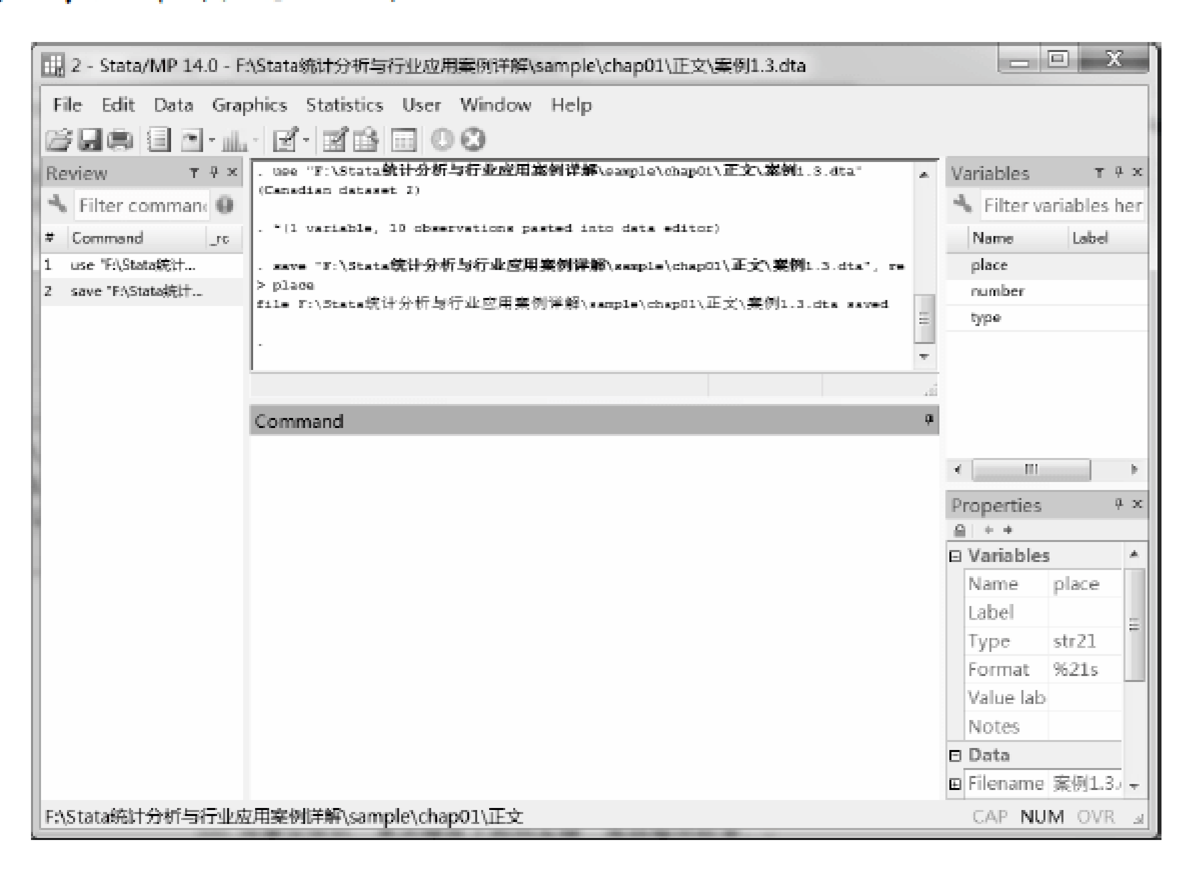


图 1.15 主界面

03 设置完毕后，按键盘上的回车键，等待输出结果。

1.4.4　结果分析

图 1.16 是生成新的分类变量来描述比赛级别的结果。

```
. tabulate type,generate(type)

  Province,
  territory
  or nation |      Freq.     Percent        Cum.
------------+-----------------------------------
   Province |          6       60.00       60.00
     Nation |          4       40.00      100.00
------------+-----------------------------------
      Total |         10      100.00
```

图 1.16　描述比赛级别的结果

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 1.17 所示的生成的分类数据“type1”和“type2”。

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 1.18 所示的生成的变量“number1”数据。该变量将“number”的取值区间划分成等宽的 3 组。图 1.18 是生成新的定序变量对场数进行定序，分到 3 个标志区间的结果。

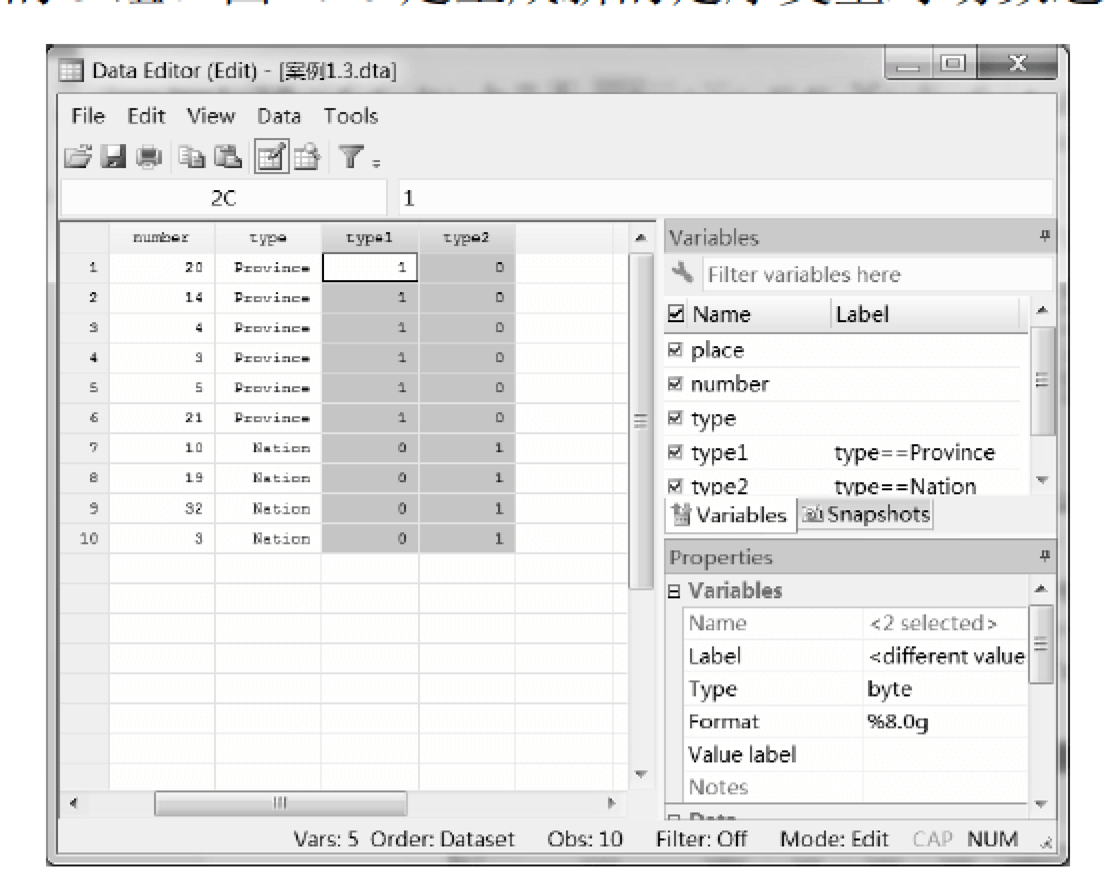


图 1.17　生成新的分类变量

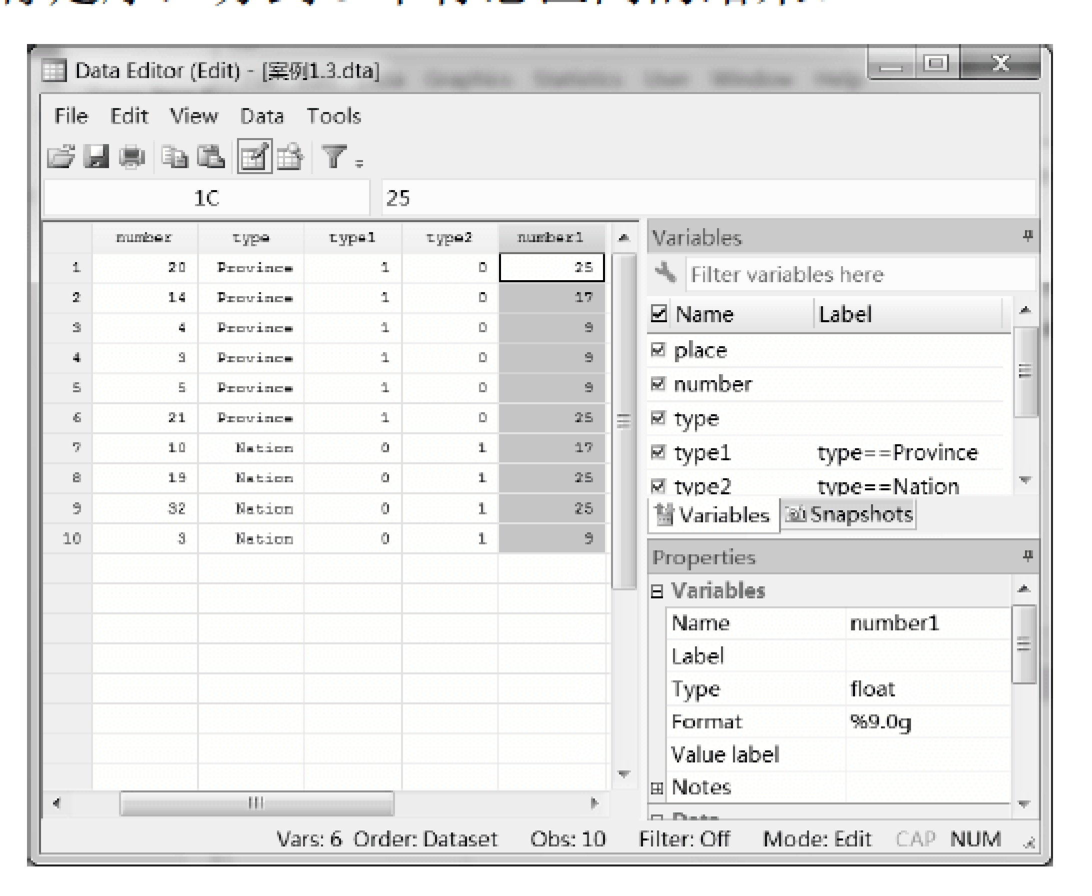


图 1.18　对场数进行定序

1.4.5　案例延伸

以本节中的案例为基础，试生成新的分类变量按数值大小对场数进行 4 类定序。

操作命令应该为：

```
sort number
generate number2=group(4)
```

在命令窗口输入命令并按回车键进行确认，选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 1.19 所示的生成的变量“number2”数据。该变量将“number”的取值按大小分成了 4 个序列。

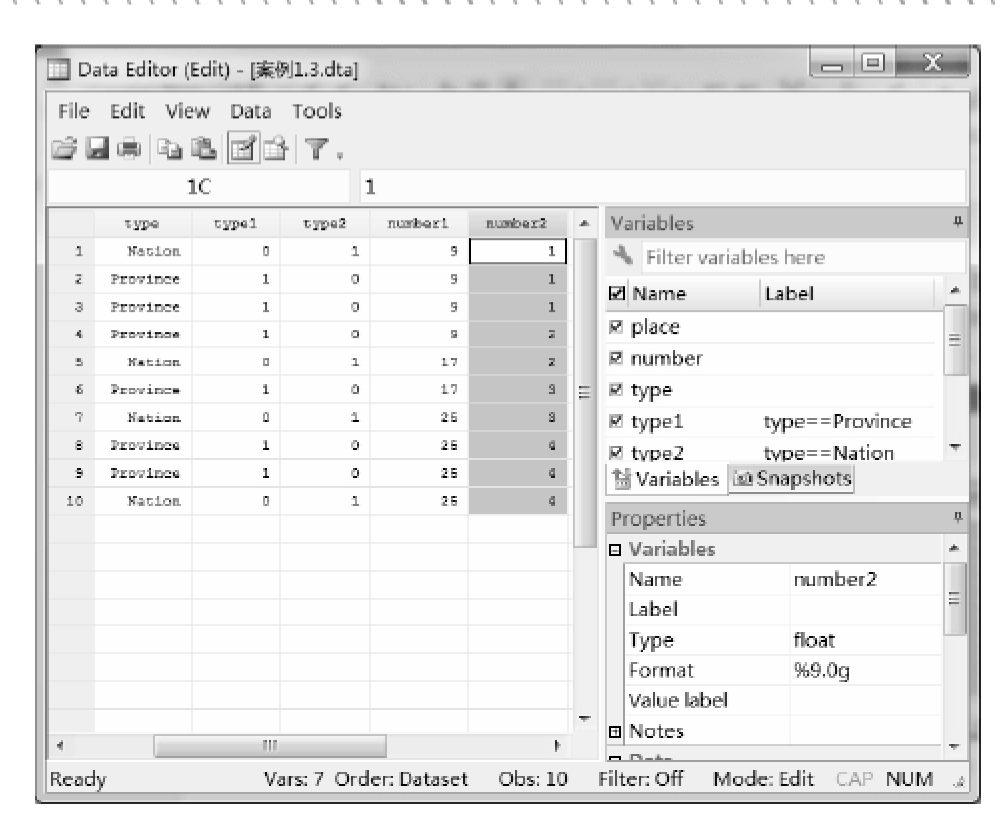


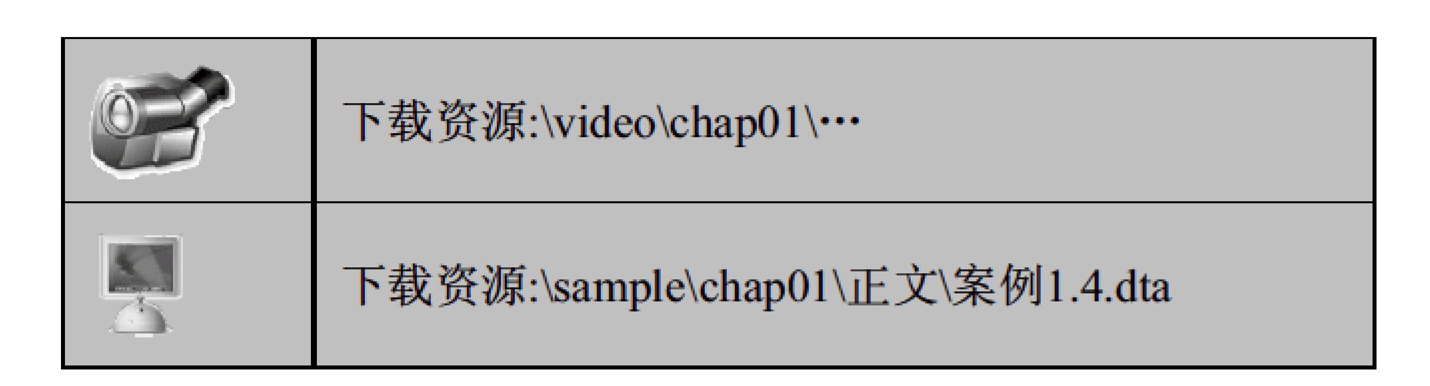
图 1.19　对场数进行 4 类定序

1.5　数据的基本操作

1.5.1　数据的基本操作概述

在对数据进行分析时，可能会遇到需要针对现有的数据进行预处理的情况。在本节中，我们将用实例讲解常用的几种处理数据的操作，包括对数据进行长短变换、把字符串数据转换成数值数据、生成随机数等。在下一节中，我们还将利用实例介绍如何定义数据子集。

1.5.2　相关数据来源


下载资源:\video\chap01\…

下载资源:\sample\chap01\正文\案例1.4.dta

【例 1.4】长江集团是一家国内大型连锁销售钢管的公司，该集团一直在北京、天津、河北、山西、内蒙古等地展开经营活动，2008—2010 年在上述地区的开店情况如表 1.6 所示。试通过操作 Stata 14.0 完成以下工作：

（1）将数据进行长短变换。

（2）将数据变换回来，并把地区字符串变量转换成数值数据。

（3）生成一个随机变量，里面包含 0~1 的 15 个随机数据。

表 1.6　长江集团在 2008—2010 年的开店情况

地区	2008年	2009年	2010年
北京	30	32	33
天津	7	8	9
河北	18	19	22
山西	60	65	32
内蒙古	26	20	15

1.5.3　Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有 4 个变量，分别是地区、2008 年店数、2009 年店数以及 2010 年店数。我们把地区变量设定为 region，把 2008 年店数变量设定为 number2008，把 2009 年店数变量设定为 number2009，把 2010 年店数变量设定为 number2010，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作在第 1.2 节中已有详细讲述。录入完成后，数据如图 1.20 所示。

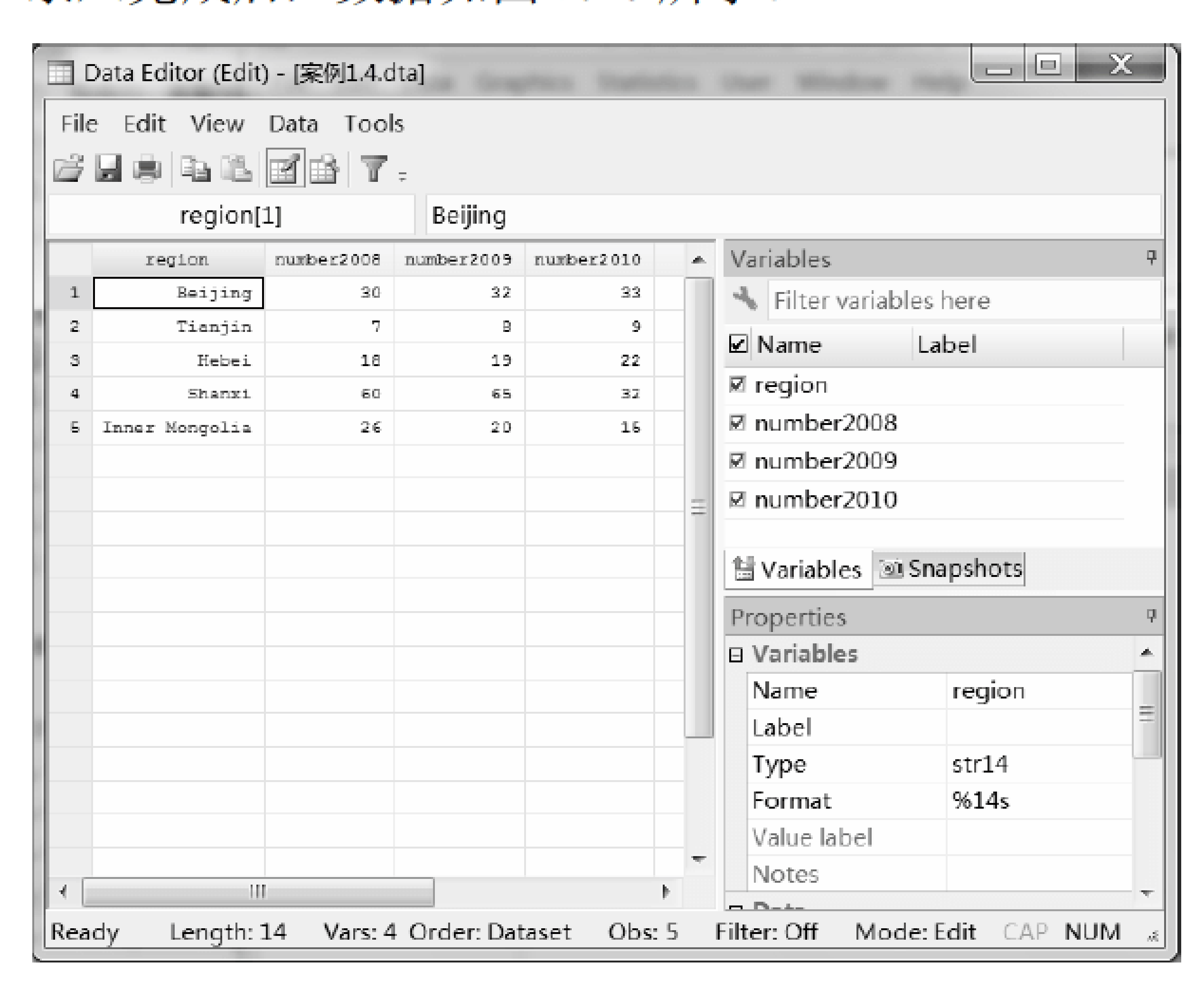


图 1.20　案例 1.4 数据

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出如图 1.21 所示的主界面。

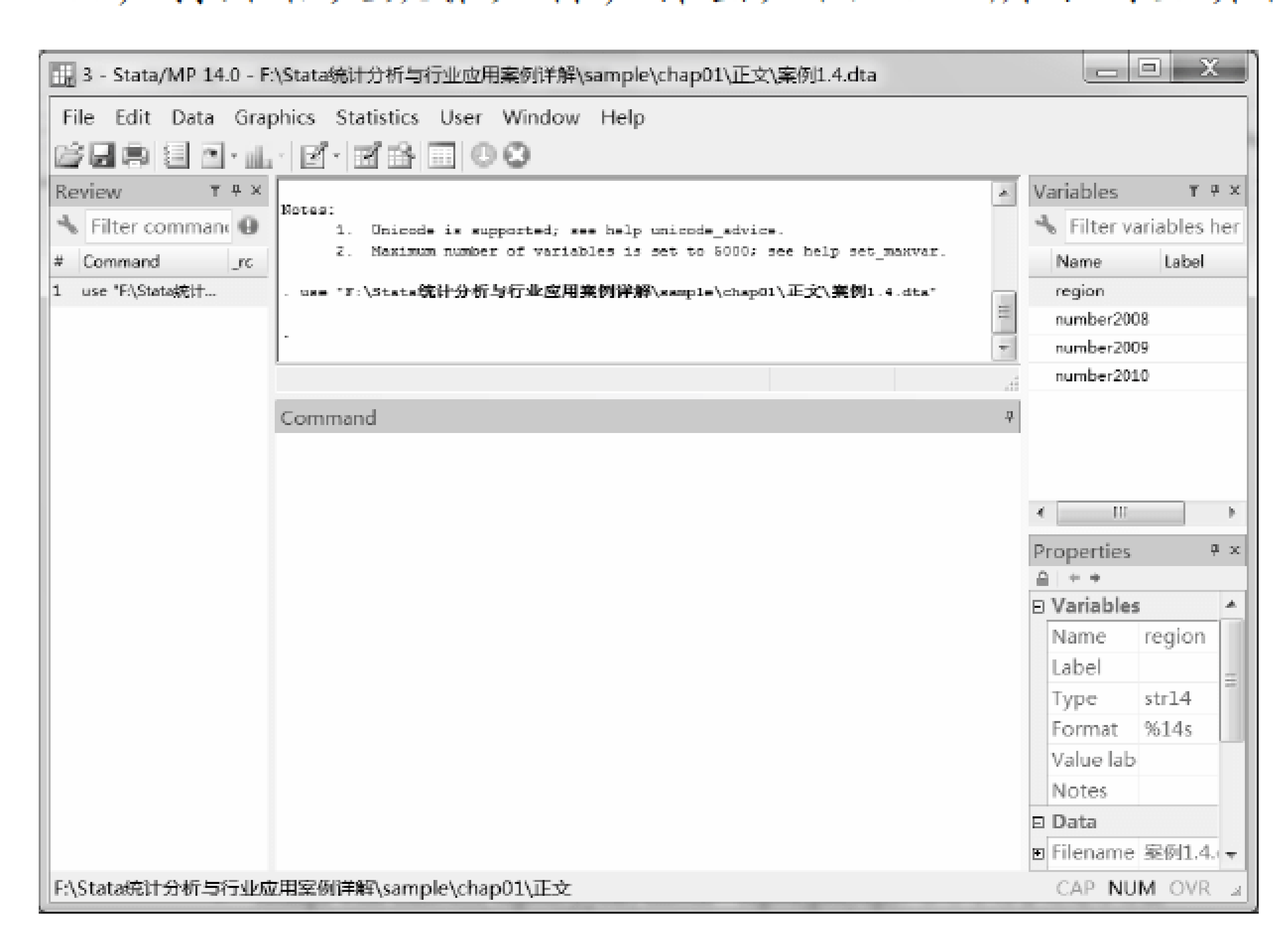


图 1.21　主界面

02 在主界面的“Command”文本框中输入操作命令并按键盘上的回车键进行确认。对应的命令分别如下：

- reshape long number,i(region) j(year)：本命令的含义是将数据进行长短变换。
- reshape wide number,i(region) j(year)。
- encode　region,gen(regi)：本命令的含义是将数据变换回来并把地区字符串变量转换成数值数据。
- Clear。
- set obs 15。
- generate suiji=uniform()：本命令的含义是生成一个随机变量，里面包含 0~1 的 15 个随机数据。

1.5.4　结果分析

图 1.22 是将数据进行长短变换的结果。

```
. reshape long number,i( region) j(year)
(note: j = 2008 2009 2010)

Data                                wide   ->   long
-----------------------------------------------------------------------------
Number of obs.                         5   ->     15
Number of variables                    4   ->      3
j variable (3 values)                      ->   year
xij variables:
       number2008 number2009 number2010    ->   number
-----------------------------------------------------------------------------
```

图 1.22　将数据进行长短变换的结果

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 1.23 所示的变换后的数据。图 1.24 是将数据变换回来并把地区字符串变量转换成数值数据的结果。

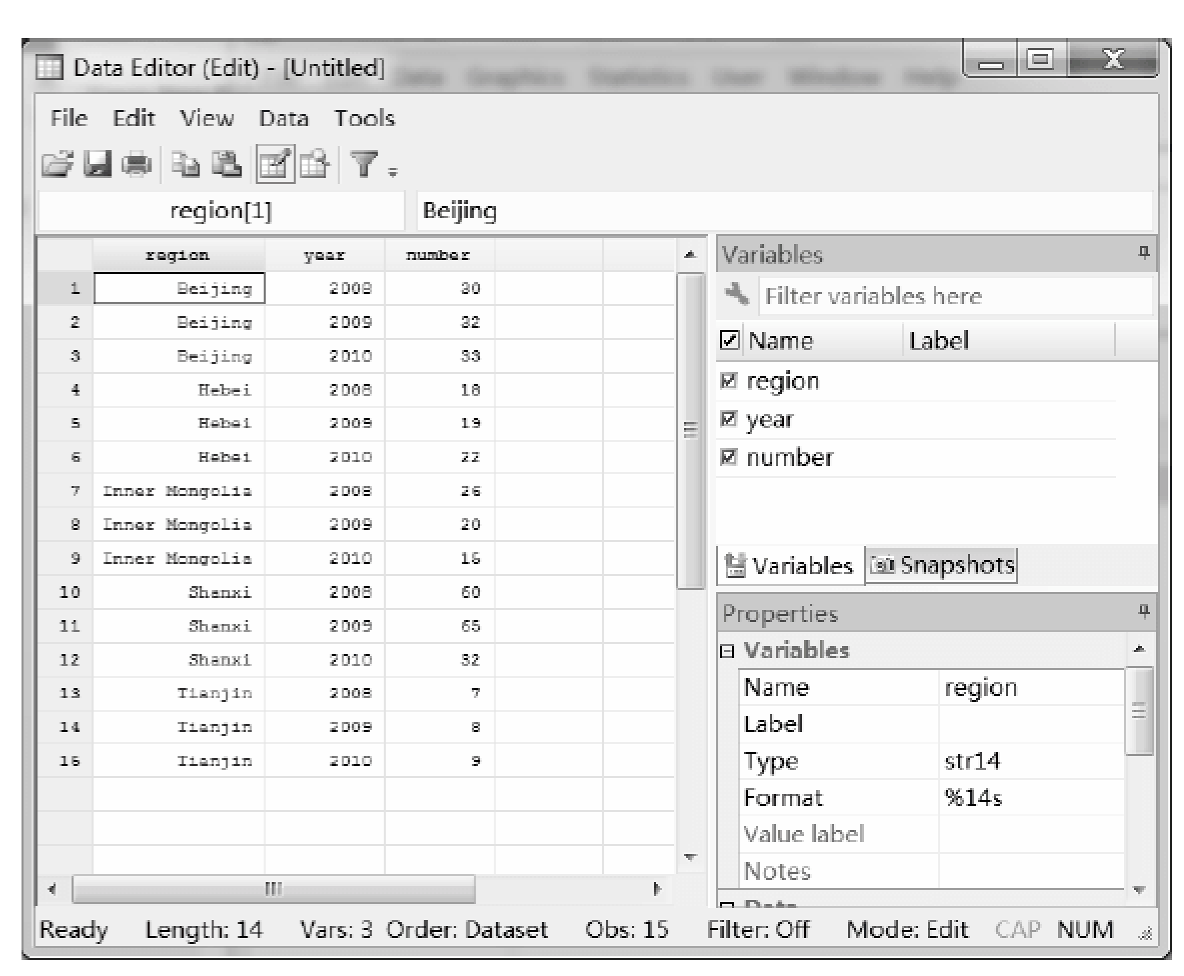


图 1.23　进行长短变换

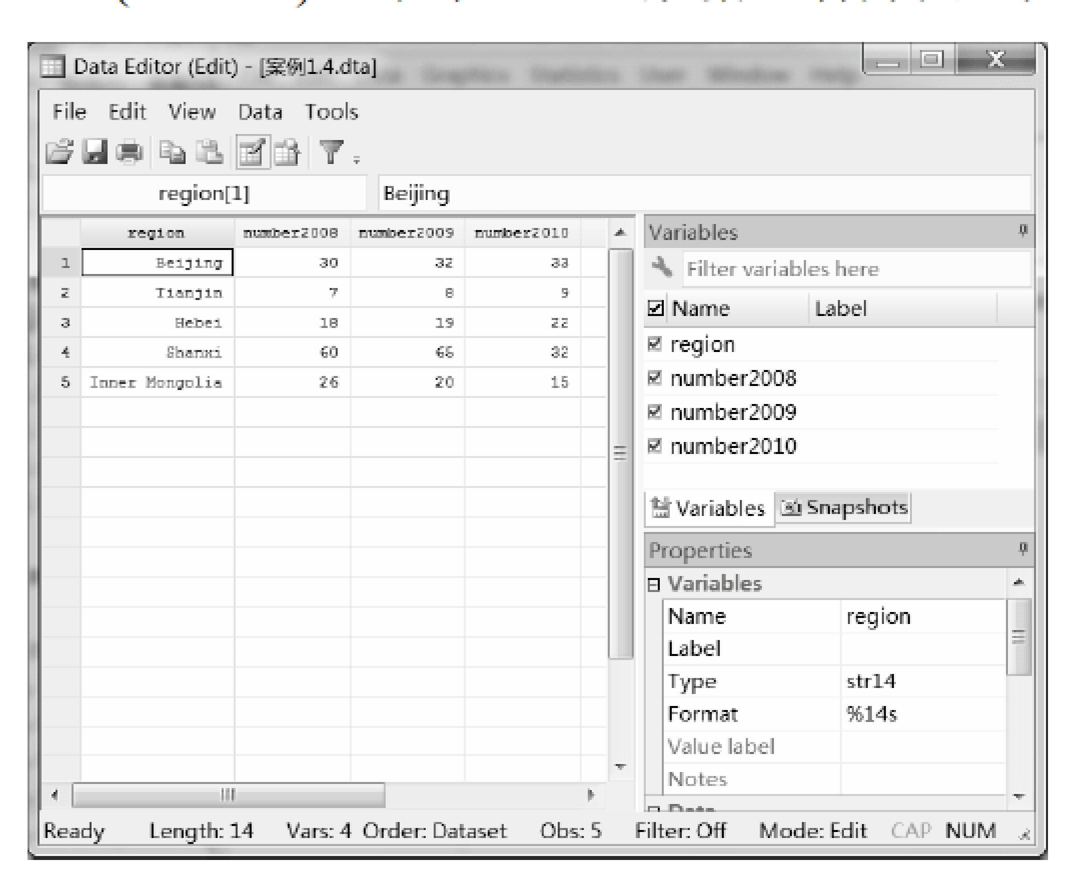
```
. reshape wide number,i( region) j(year)
(note: j = 2008 2009 2010)

Data                               long   ->   wide
-----------------------------------------------------------------------------
Number of obs.                       15   ->      5
Number of variables                   3   ->      4
j variable (3 values)              year   ->   (dropped)
xij variables:
                                 number   ->   number2008 number2009 number2010
-----------------------------------------------------------------------------
```

图 1.24　转换成数值数据的结果

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 1.25 所示的变换后的数据。

在将数据变换回来以后，输入第 2 条命令，通过选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，如图 1.26 所示。



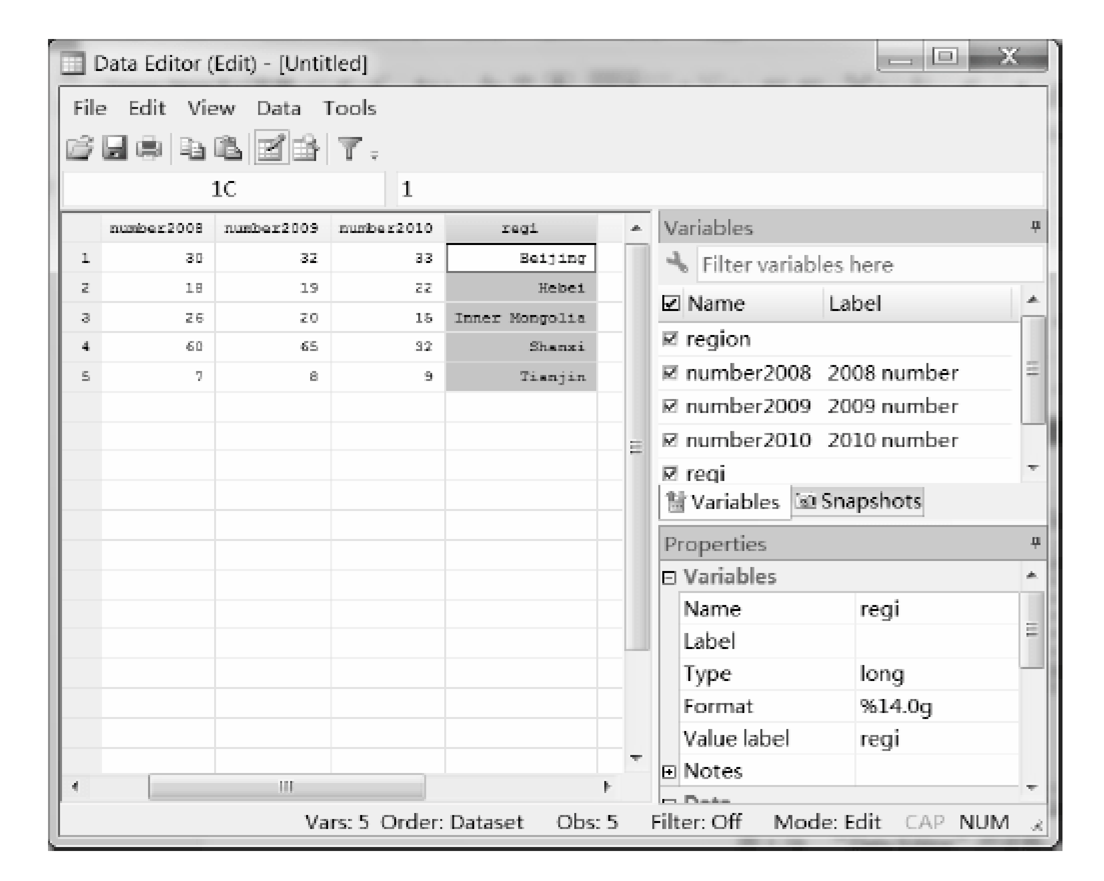


图 1.25　变换后的数据　　　　图 1.26　查看数据

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 1.27 所示的生成后的随机数据。

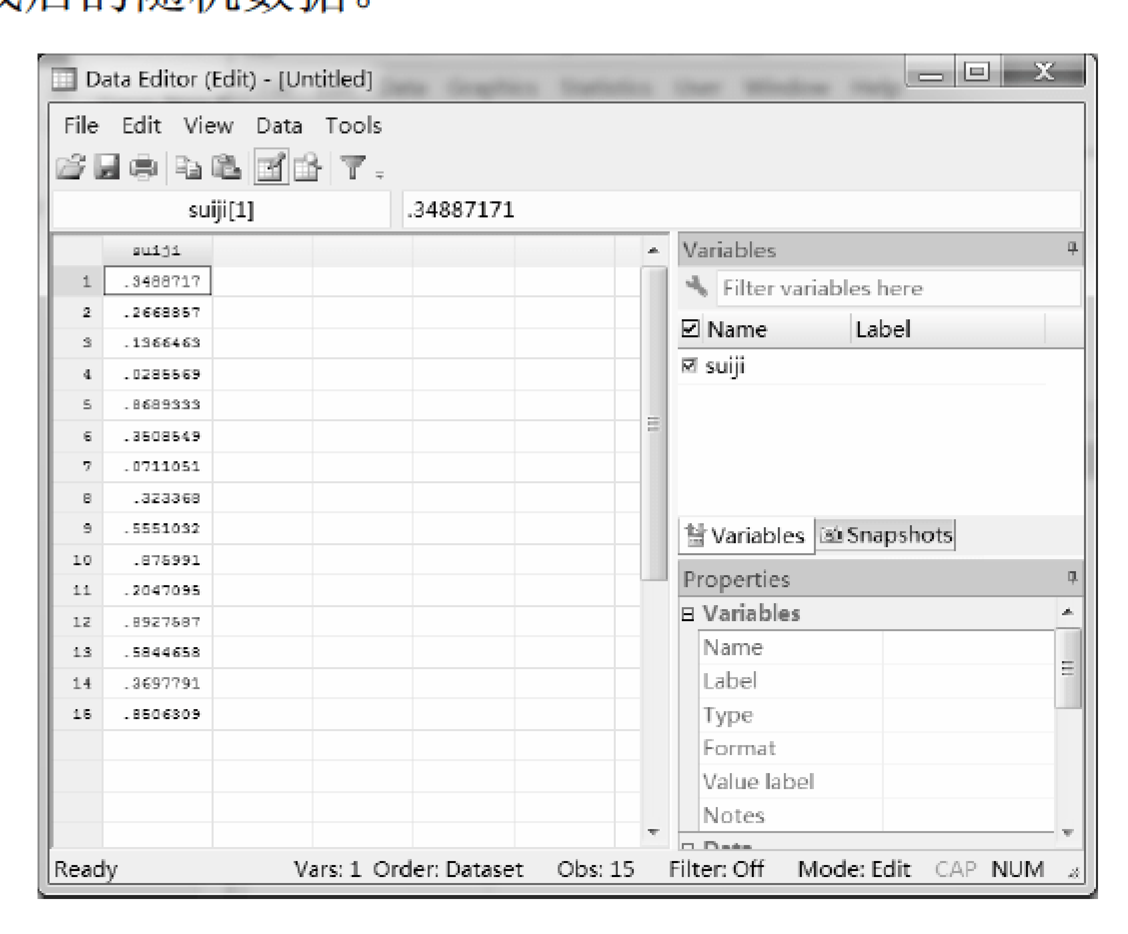


图 1.27　随机数据

1.5.5 案例延伸

在定义随机数据时，系统命令默认的区间范围即是[0,1]，那么如何实现自由取值呢？例如，从[9,18]之间随机取出 15 个数据。

操作命令应该相应地修改为如下形式：

```
clear
set obs 15
generate suiji=9+9*uniform()
```

在命令窗口输入命令并按回车键进行确认的结果如图 1.28 所示。

那么如何选取整数呢？

操作命令应该相应地修改为如下形式：

```
clear
set obs 15
generate suiji=9+trunc(9*uniform())
```

在命令窗口输入命令并按回车键进行确认的结果如图 1.29 所示。

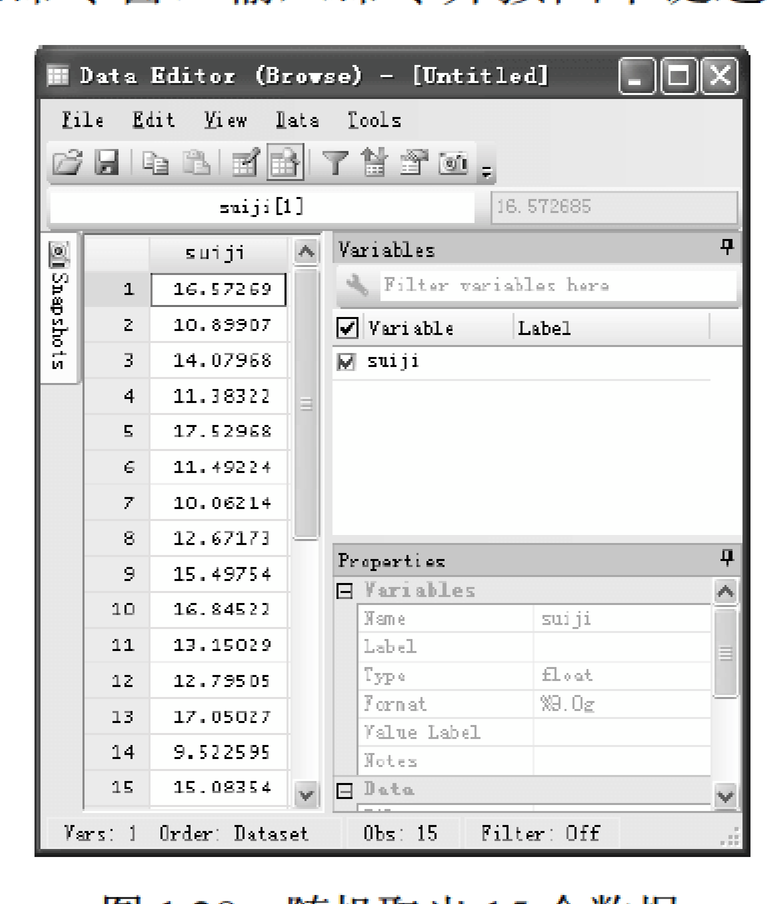


图 1.28 随机取出 15 个数据

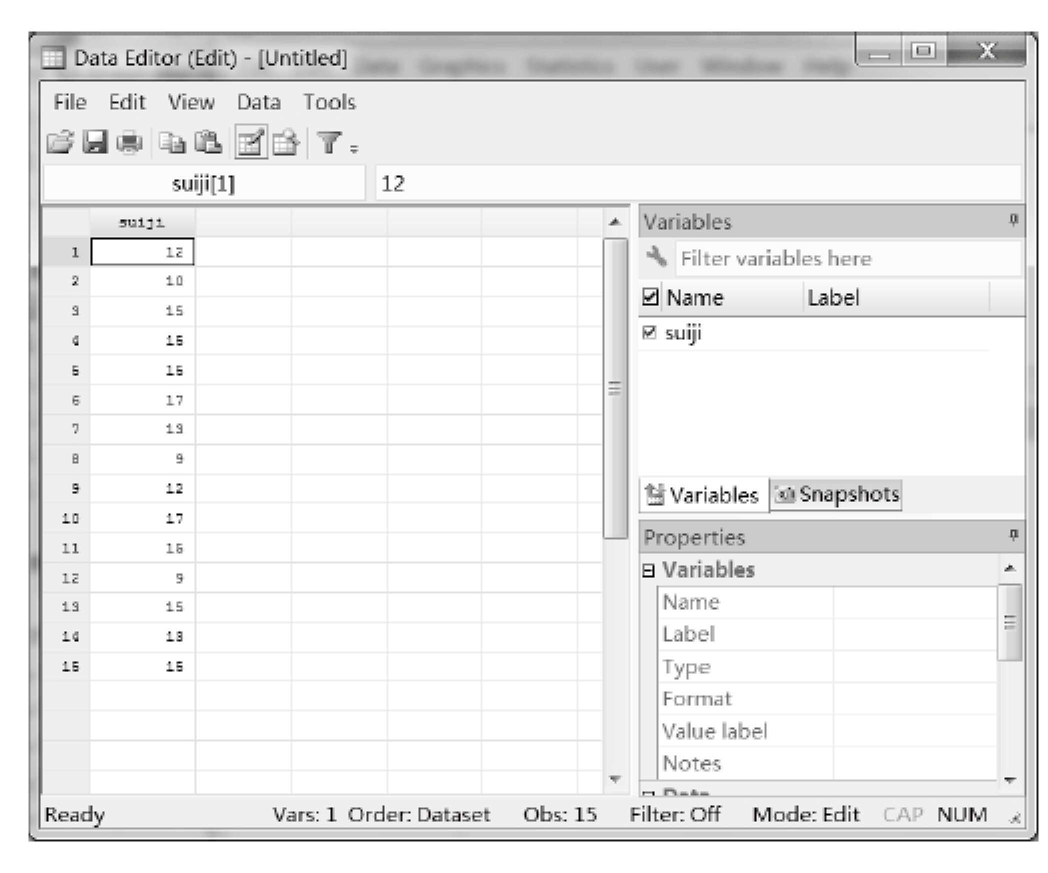


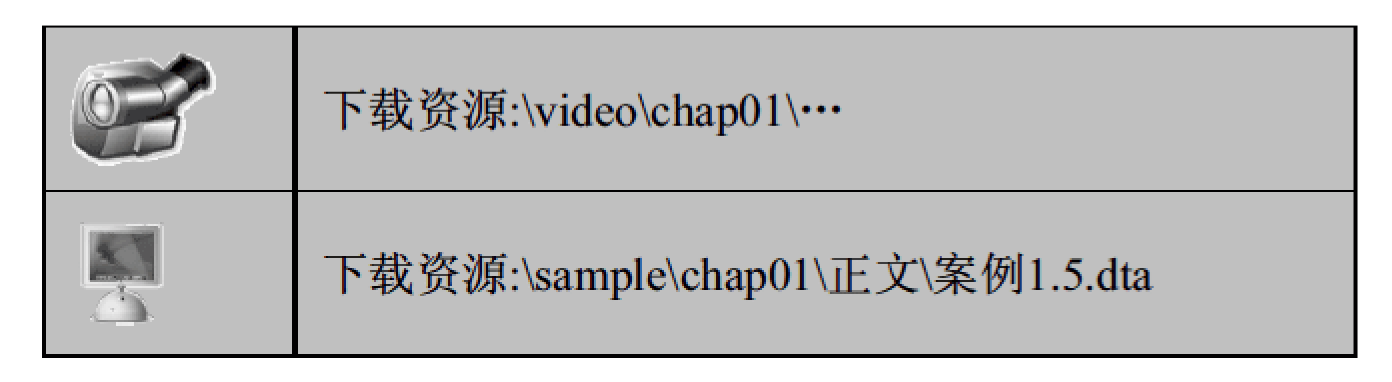
图 1.29 取整

1.6 定义数据的子集

1.6.1 定义数据的子集概述

在很多情况下，现有的 Stata 数据达不到分析要求，我们需要截取出数据的一部分进行分析，或者删除不需要进入分析范围的数据，这时我们就需要用到 Stata 的定义数据子集功能。在本节中，我们将通过实例的方式讲述定义数据子集的基本操作。

1.6.2　相关数据来源



	下载资源:\video\chap01\…
	下载资源:\sample\chap01\正文\案例1.5.dta

【例 1.5】试通过操作案例 1.5.dta 完成以下工作。

（1）列出第 3 条数据。

（2）列出第 1~3 条数据。

（3）列出变量值“shangjiao”最小的两条数据。

（4）列出变量值“year”大于 2005 的数据。

（5）列出变量值“year”大于 2007 且变量值“shangjiao”大于 865 的数据。

（6）删除第 3 条数据。

（7）删除变量值“year”等于 2005 的数据。

（8）删除变量值“year”大于 2005 且变量值“shangjiao”大于 865 的数据。

1.6.3　Stata 分析过程

分析步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出如图 1.30 所示的主界面。

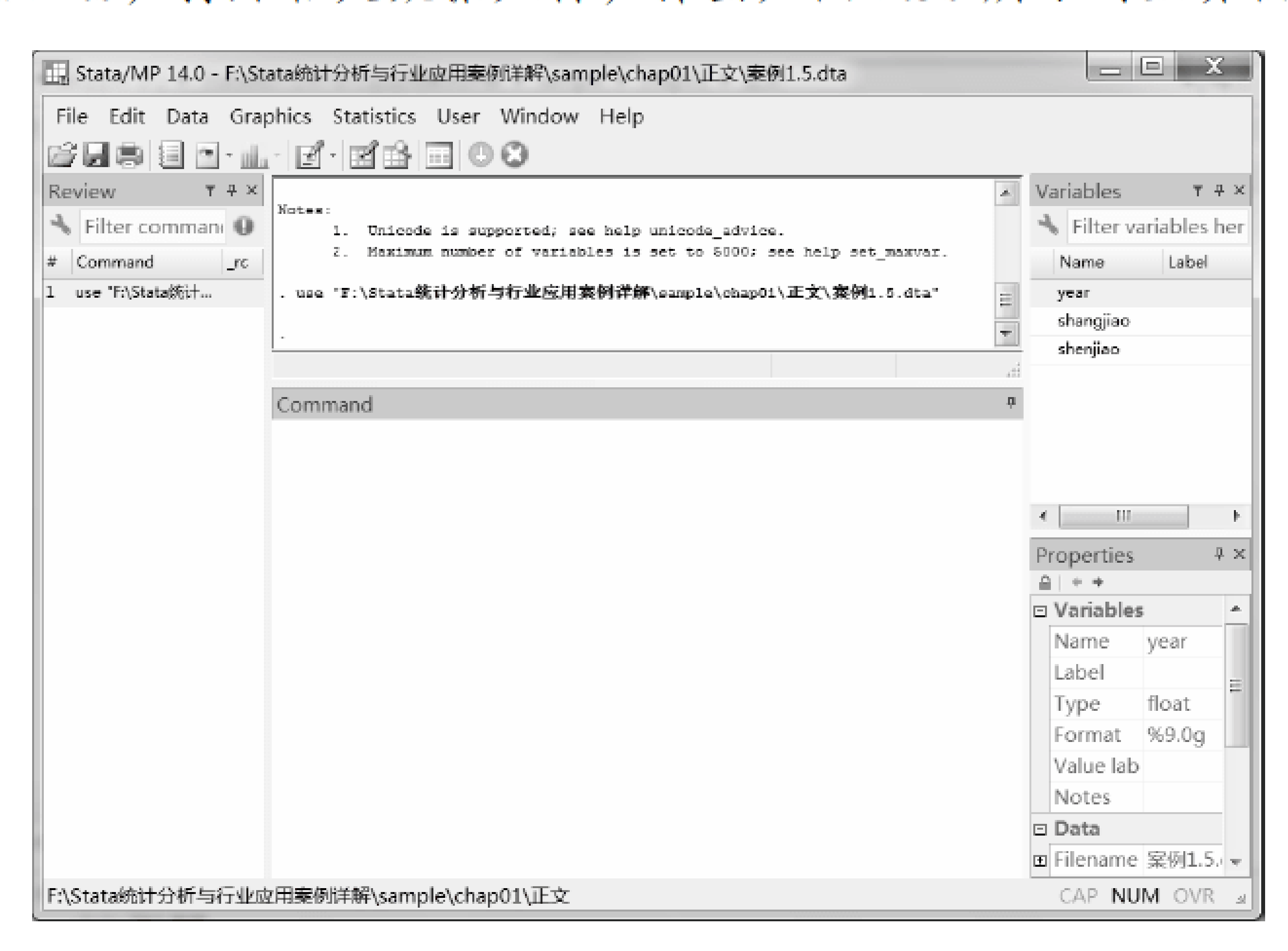


图 1.30　主界面

02 在主界面的“Command”文本框中输入操作命令并按键盘上的回车键进行确认。对应的命令如下。

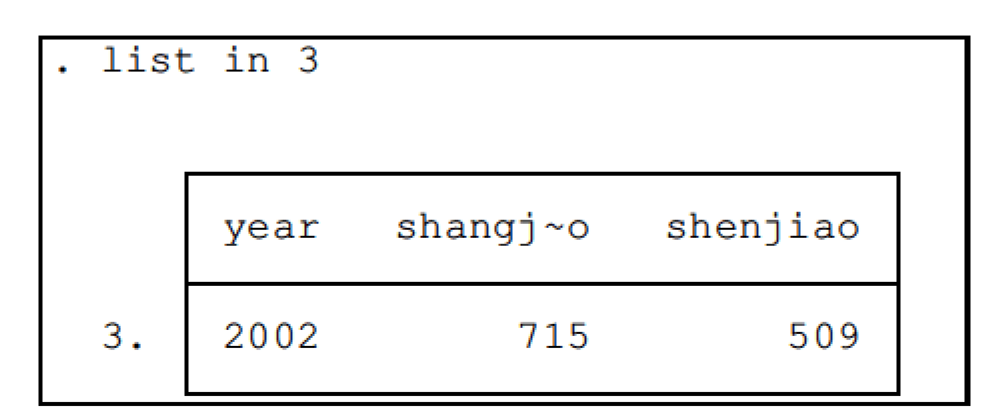
- list in 3：本命令的含义是列出第 3 条数据。
- list in 1/3：本命令的含义是列出第 1~3 条数据。

- sort shangjiao list year shangjiao shenjiao in 1/2: 本命令的含义是列出变量值“shangjiao”最小的两条数据。
- list if year>2005：本命令的含义是列出变量值“year”大于 2005 的数据。
- list if year>2007 & shangjiao>865：本命令的含义是列出变量值“year”大于 2007 且变量值“shangjiao”大于 865 的数据。
- drop in 3：本命令的含义是删除第 3 条数据。
- drop if year==2005：本命令的含义是删除变量值“year”等于 2005 的数据。
- drop if year>2005 & shangjiao>865：本命令的含义是删除变量值“year”大于 2005 且变量值“shangjiao”大于 865 的数据。

1.6.4 结果分析

图 1.31 是列出第 3 条数据的结果。

图 1.32 是列出第 1~3 条数据的结果。

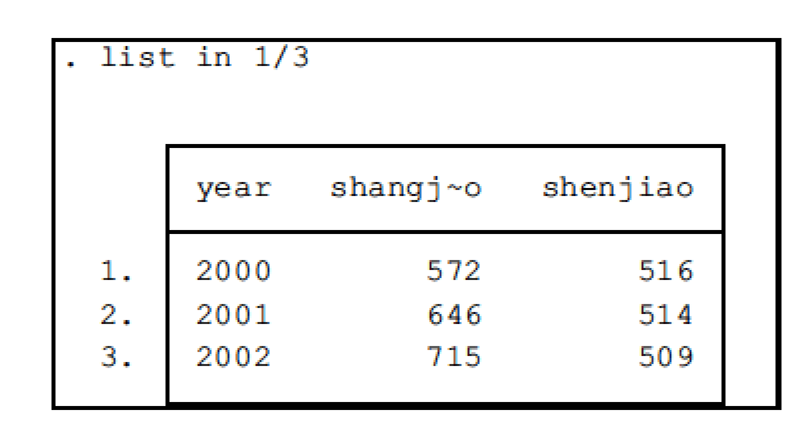


```
. list in 3

     year   shangj~o   shenjiao

 3.  2002        715        509
```

图 1.31 分析结果 1



```
. list in 1/3

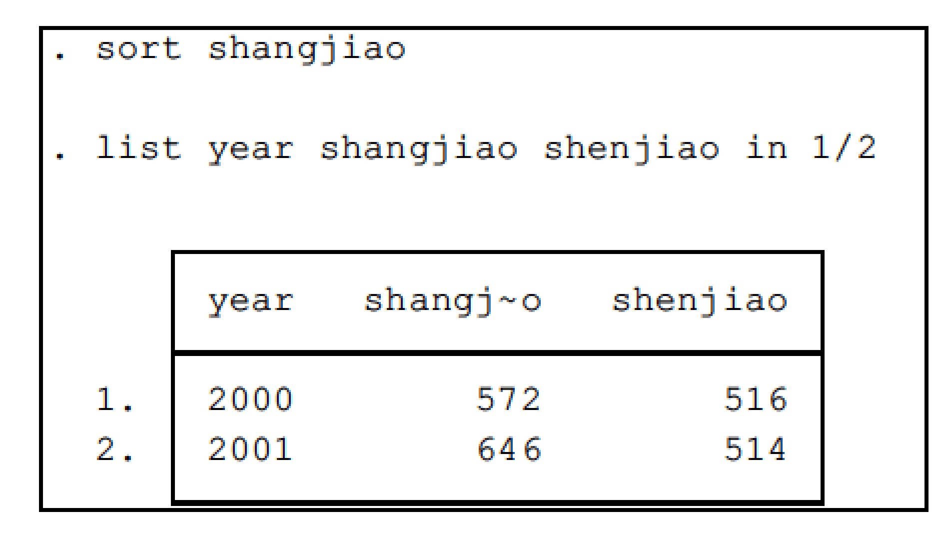
     year   shangj~o   shenjiao

 1.  2000        572        516
 2.  2001        646        514
 3.  2002        715        509
```

图 1.32 分析结果 2

图 1.33 是列出变量值“shangjiao”最小的两条数据结果。

图 1.34 是列出变量值“year”大于 2005 的数据结果。



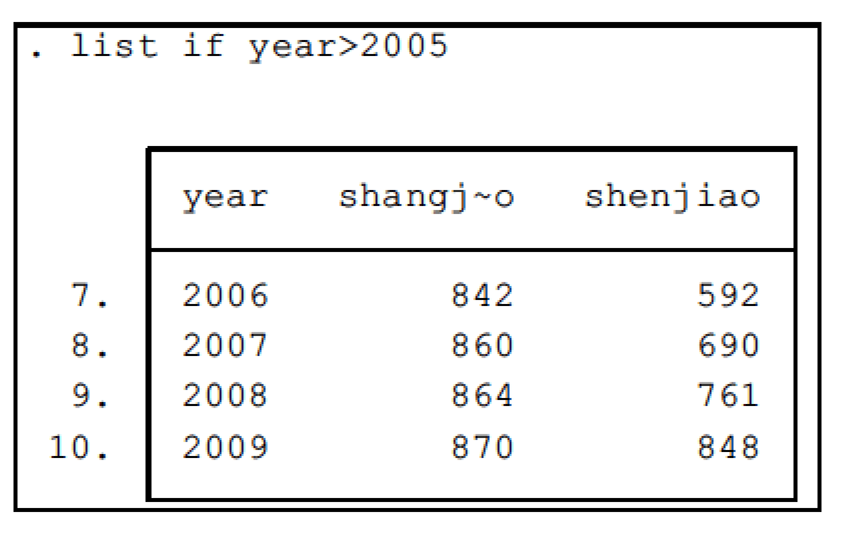
```
. sort shangjiao

. list year shangjiao shenjiao in 1/2

     year   shangj~o   shenjiao

 1.  2000        572        516
 2.  2001        646        514
```

图 1.33 分析结果 3



```
. list if year>2005

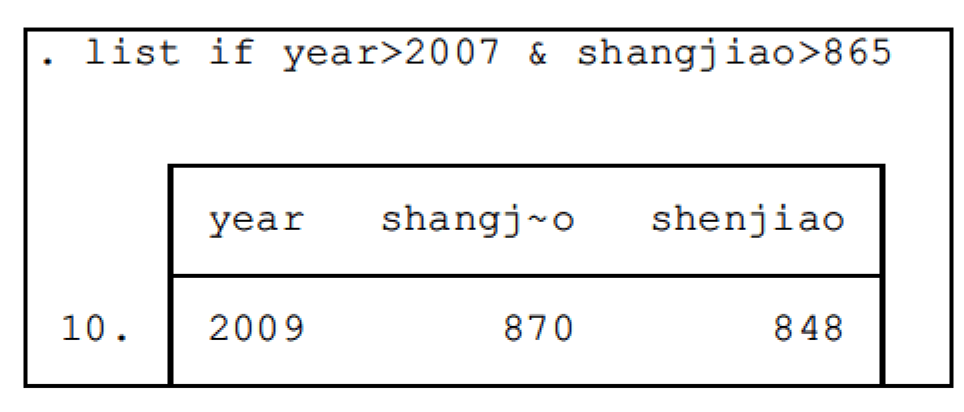
      year   shangj~o   shenjiao

  7.  2006        842        592
  8.  2007        860        690
  9.  2008        864        761
 10.  2009        870        848
```

图 1.34 分析结果 4

图 1.35 是列出变量值“year”大于 2007 且变量值“shangjiao”大于 865 的数据结果。

图1.36是删除第3条数据的结果。



```
. list if year>2007 & shangjiao>865

      year   shangj~o   shenjiao

 10.  2009        870        848
```

图 1.35 分析结果 5

```
. drop in 3
(1 observation deleted)
```

图 1.36 分析结果 6

图 1.37 是删除变量值“year”等于 2005 的数据结果。

图 1.38 是删除变量值“year”大于 2005 且变量值“shangjiao”大于 865 的数据结果。

```
. drop if year==2005
(1 observation deleted)
```

图 1.37　分析结果 7

```
. drop if year>2005 & shangjiao>865
(1 observation deleted)
```

图 1.38　分析结果 8

1.6.5　案例延伸

我们在上述的 Stata 命令中用到了 Stata 中的关系运算符和逻辑运算符。Stata 14.0 中共支持 6 种关系运算符和 3 种逻辑运算符，如表 1.7 和表 1.8 所示。

表 1.7　关系运算符

关系运算符	含义	关系运算符	含义	关系运算符	含义
==	等于	! =	不等于	>	大于
<	小于	>=	大于等于	<=	小于等于

表 1.8　逻辑运算符

逻辑运算符	含义	逻辑运算符	含义	逻辑运算符	含义
&	与	|	或	!	非

1.7　本章习题

（1）表 1.9 记录的是两家公司近些年的招聘员工数据。试创建 Stata 格式的数据文件并保存。

表 1.9　两家公司近些年的招聘员工数据

年份	X公司	Y公司
2000	45	58
2001	66	77
2002	38	44
2003	22	22
2004	58	34
2005	33	57
2006	44	52
2007	86	69
2008	102	61
2009	41	84

（2）某连锁公司在全国各地区的销售人员数量以及销售总额数据如表 1.10 所示。请使用 Stata 命令进行操作：①试生成新的变量来描述各地区的人均销售额情况；②试生成人均销售额变量来替代原有的销售总额变量；③对生成的人均销售额变量数据均做除以 10 的处理；④对销售人员数量变量进行对数平滑处理，从而产生新的变量。

表 1.10　某连锁公司在全国各地区的销售人员数量以及销售总额数据

地区	销售人员数量/人	销售总额/万元
北京	50	250 000
天津	30	90 000
河北	50	300 000
山西	60	420 000
内蒙古	40	180 000
…	…	…
青海	40	80 000
宁夏	20	20 000
新疆	25	37 500

（3）某当红歌星近两年来在各地举办演唱会的情况如表 1.11 所示。试使用 Stata 14.0 对数据进行以下操作：①试生成新的分类变量来描述演唱会类型；②试生成新的定序变量对场数进行定序，分到 3 个标志区间。③试生成新的分类变量，按数值大小对场数进行 4 类定序。

表 1.11　某当红歌星最近两年来在各地举行演唱会情况

地点	场数	演唱会级别
北京	17	中型
浙江	16	中型
天津	5	中型
福建	3	中型
江苏	5	中型
山东	23	中型
美国	12	大型
日本	17	大型
韩国	32	大型
新加坡	5	大型

（4）某足球俱乐部以培养优秀年轻球员而出名，当红的 5 名明星队员在 2008—2010 年赛季的进球情况如表 1.12 所示。试通过操作 Stata 14.0 完成以下工作：

①将数据进行长短变换。

②将数据变换回来，并把球员名称字符串变量转换成数值数据。

③生成一个随机变量，里面包含 0~1 的 15 个随机数据。

表 1.12　某足球俱乐部的 5 名明星队员在 2008—2010 年赛季的进球情况

球员名称	2008年	2009年	2010年
a	35	32	36
b	9	7	19
c	28	19	22
d	61	55	22
e	26	22	15

（5）试通过操作案例 1.5.dta 完成以下工作：

①列出第 3 条数据。

②列出第 1~3 条数据。
③列出变量值“shenjiao”最小的两条数据。
④列出变量值“year”大于 2003 的数据。
⑤列出变量值“year”大于 2003 且变量值“shenjiao”大于 55 的数据。
⑥删除第 3 条数据。
⑦删除变量值“year”等于 2003 的数据。
⑧删除变量值“year”大于 2004 且变量值“shenjiao”大于 50 的数据。

第 2 章　Stata 图形绘制

众所周知，图形是对数据分析结果以及其他综合分析一种很好的展示方式。制图功能一直是Stata 的强项，也是许多软件使用者选择该软件进行数据分析的重要理由之一。经过 Stata 公司编程人员的长期不懈努力，制图功能在 Stata 14.0 版本中已经非常完善，比较以前的版本，不仅形成图形的能力得到增强，图形输出的外观和选择也得到了大大改进。限于篇幅，本章将介绍用户最常用的几种绘图功能。软件使用者常用的制图功能有直方图、散点图、曲线标绘图、连线标绘图、箱图、饼图、条形图、点图等。下面我们一一介绍这几种制图功能在实例中的应用。

2.1　实例一——直方图

2.1.1　直方图的功能与意义

直方图（Histogram）又称柱状图，是一种统计报告图，由一系列高度不等的纵向条纹或线段表示数据分布的情况。一般用横轴表示数据类型，纵轴表示分布情况。通过绘制直方图，可以较为直观地传递有关数据的变化信息，使数据使用者能够较好地观察数据波动的状态，使数据决策者能够依据分析结果确定在什么地方需要集中力量改进工作。

2.1.2　相关数据来源

	下载资源:\video\chap02\…
	下载资源:\sample\chap02\正文\案例2.1.dta

【例 2.1】为了解我国各地区技工学校的建设情况，某课题组搜集整理了 2009 年我国 29 个省市的技工学校数量的数据，如表 2.1 所示。试通过绘制直方图来直观地反映我国技工学校的建设情况。

表 2.1　2009 年我国 29 个省市技工学校的数量

地区	技工学校个数
北京	38
天津	44
河北	164
山西	109
内蒙古	32
…	…

（续表）

地区	技工学校个数
青海	18
宁夏	20
新疆	60

2.1.3　Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有两个变量，分别是地区和数量。我们把地区变量设定为 region，把数量变量设定为 number，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作我们在第 1 章中已有详细讲述。录入完成后，数据如图 2.1 所示。

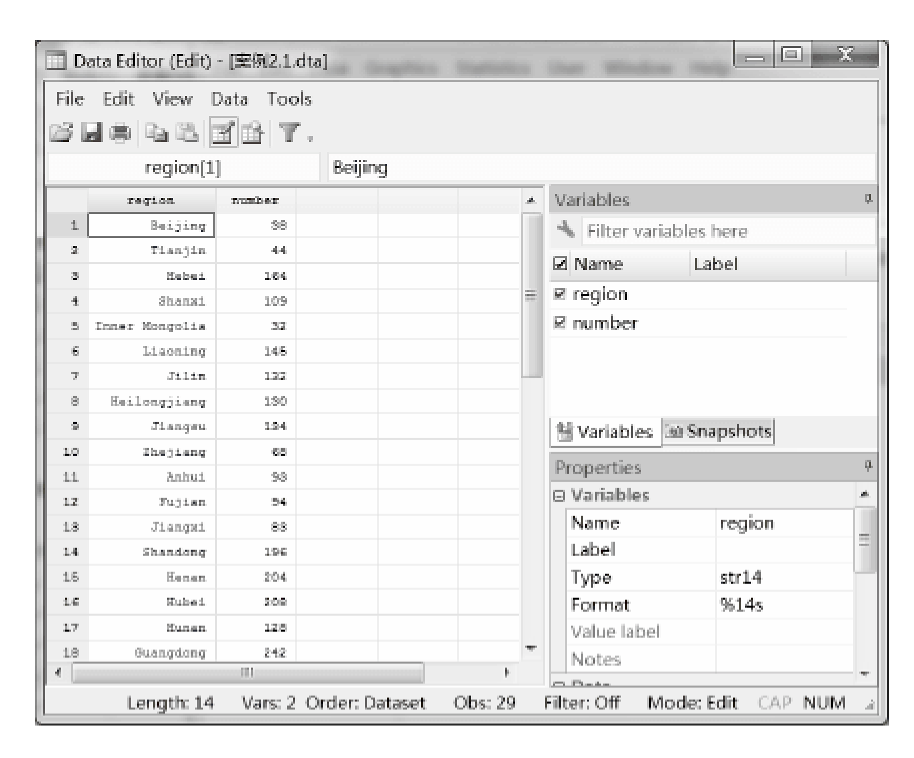


图 2.1　案例 2.1 数据

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入命令：histogram number,Frequency。

03 设置完毕后，按键盘上的回车键，等待输出结果。

2.1.4　结果分析

上述操作结束后，Stata 14.0 将弹出如图 2.2 所示的直方图。

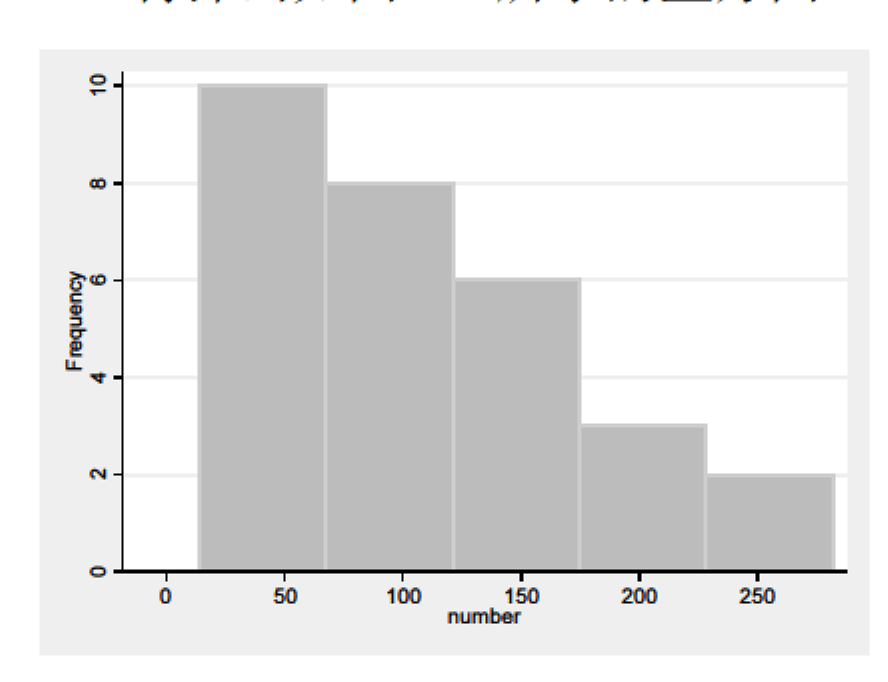


图 2.2　直方图 1

通过观察直方图，可以比较轻松地看出我国的技工学校建设情况，某省市拥有技工学校的数量和与之处于同一区间省市的数量是负相关的，也就是说，拥有技工学校数量较多的省市较少，拥有技工学校数量较少的省市较多。

2.1.5 案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

1. 延伸 1：给图形增加标题

例如，我们要给图形增加标题的名称“案例 2.1 结果”，那么操作命令就应该相应地修改为：

```
histogram number,frequency title("案例 2.1 结果")
```

在命令窗口输入命令并按回车键进行确认，结果如图 2.3 所示。

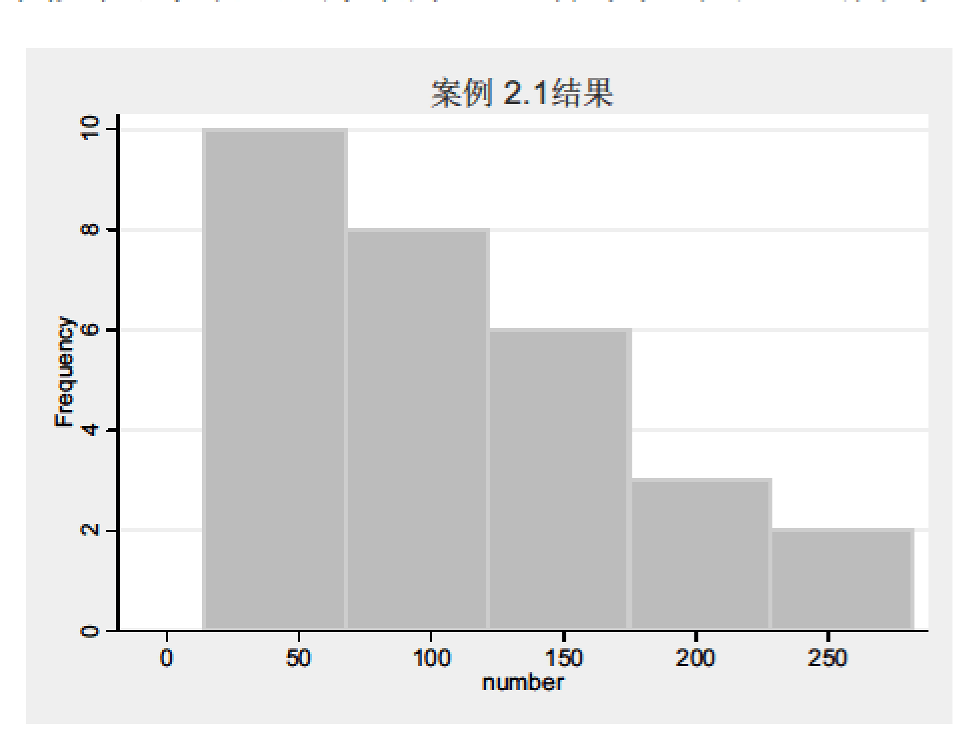


图 2.3 直方图 2

2. 延伸 2：给坐标轴增加数值标签并设定间距

例如，我们要在延伸 1 的基础上对 X 轴添加数值标签，取值为 0~300，间距为 25，对 Y 轴添加数值标签，取值为 0~10，间距为 1，那么操作命令就应该相应地修改为：

```
histogram number,frequency title("案例 2.1 结果")xlabel(0(25)300) ylabel(0(1)10)
```

在命令窗口输入命令并按回车键进行确认，结果如图 2.4 所示。

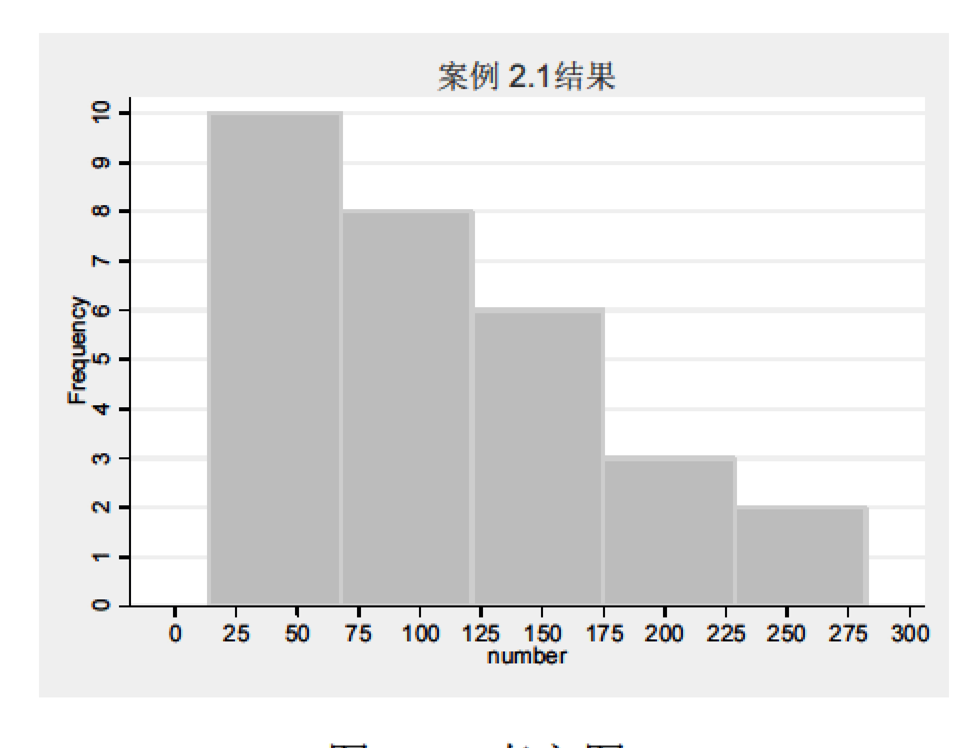


图 2.4 直方图 3

3．延伸 3：显示坐标轴的刻度

例如，我们要在延伸 2 的基础上对 Y 轴添加刻度，取值为 0~10，间距为 0.5，那么操作命令就应该相应地修改为：

```
histogram number,frequency title("案例 2.1 结果")
xlabel(0(25)300) ylabel(0(1)10) ytick(0(0.5)10)
```

在命令窗口输入命令并按回车键进行确认，结果如图 2.5 所示。

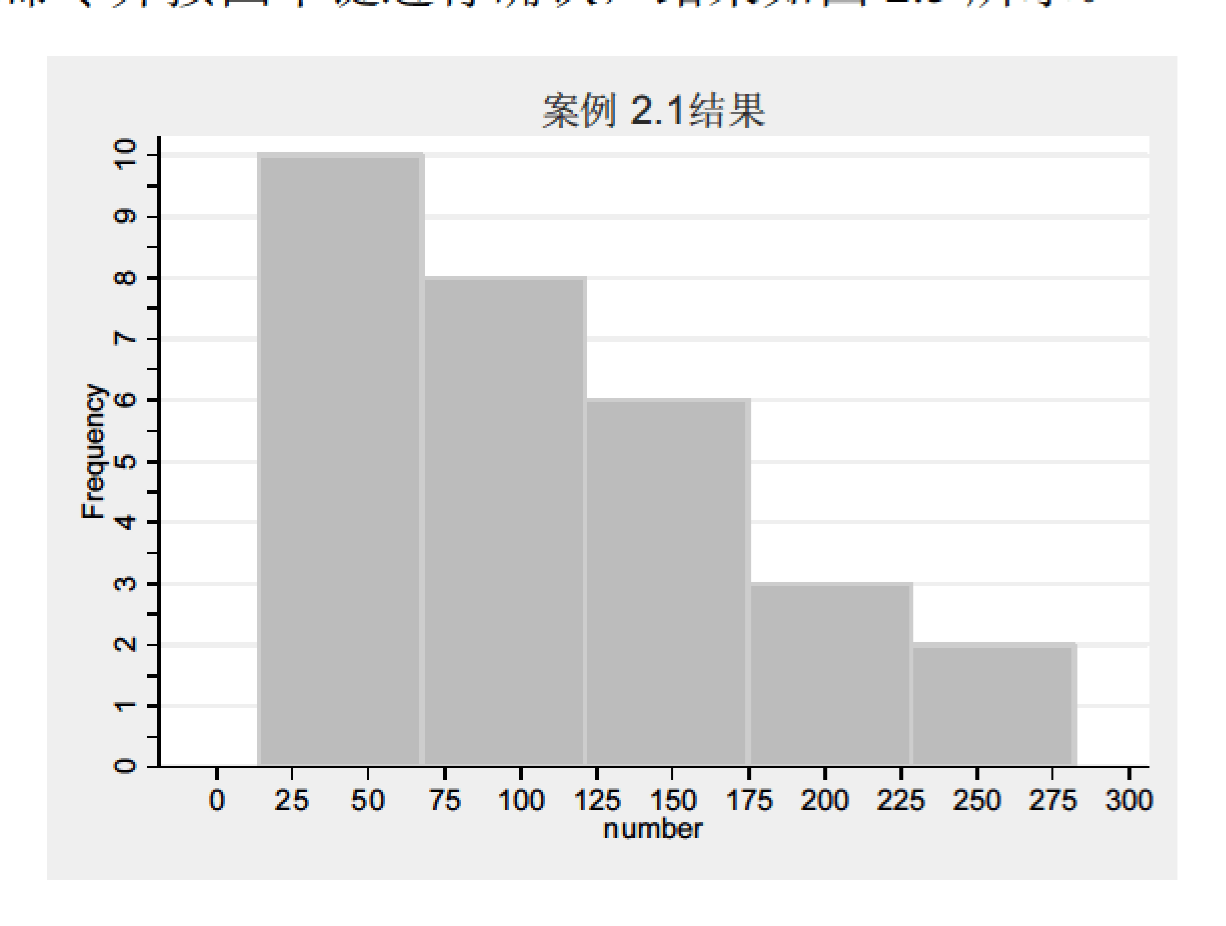


图 2.5　直方图 4

4．延伸 4：设定直方图的起始值以及直方条的宽度

例如，我们要在延伸 3 的基础上进行改进，使直方图的第 1 个直方条从 10 开始，每一个直方条的宽度为 25，那么操作命令就应该相应地修改为：

```
histogram number,frequency title("案例 2.1 结果")
xlabel(0(25)300) ylabel(0(1)10) ytick(0(0.5)10) start(10) width(25)
```

在命令窗口输入命令并按回车键进行确认，结果如图 2.6 所示。

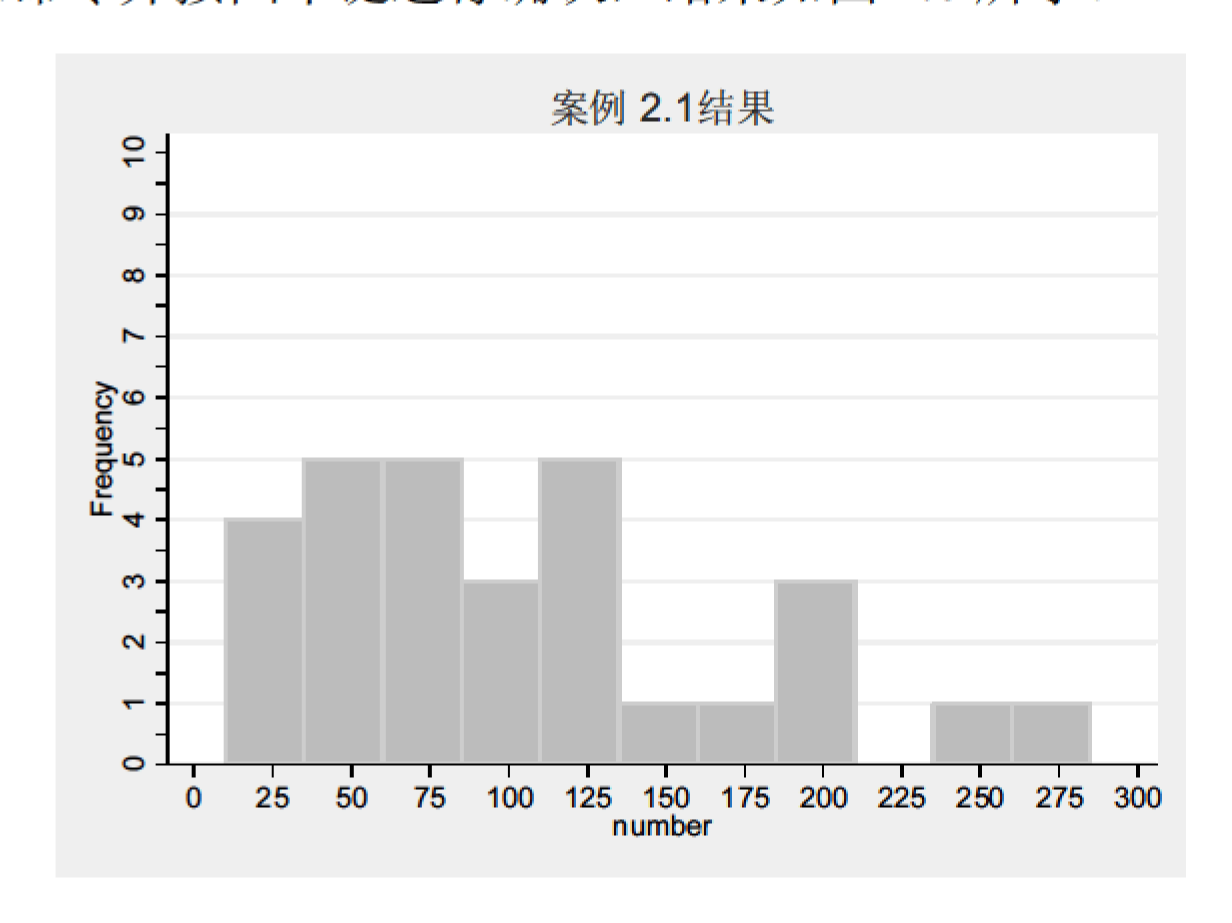


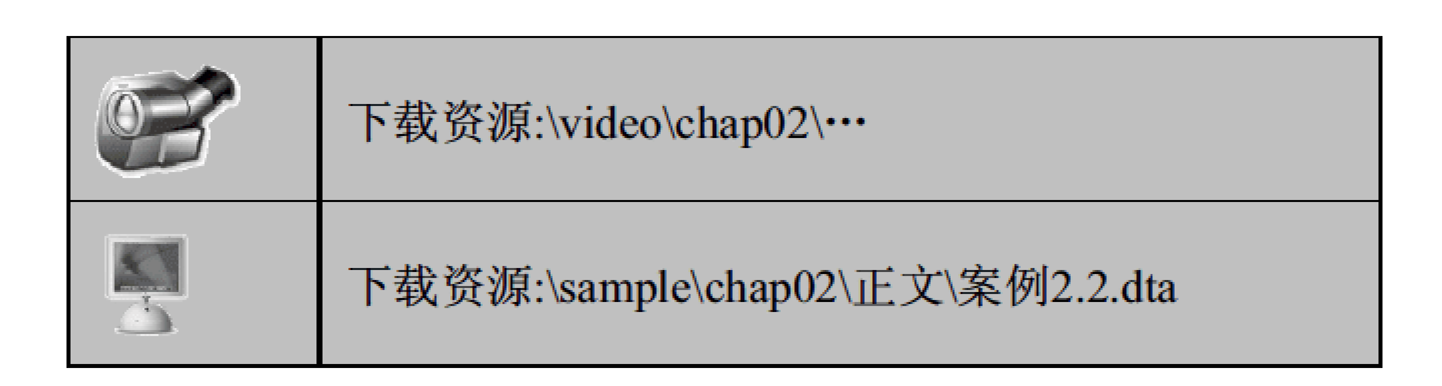
图 2.6　直方图 5

2.2 实例二——散点图

2.2.1 散点图的功能与意义

作为对数据进行预处理的重要工具之一，散点图（Scatter Diagram）功能深受专家、学者们的喜爱。散点图的简要定义就是点在直角坐标系平面上的分布图。研究者对数据制作散点图的主要出发点是通过绘制该图来观察某变量随另一变量变化的大致趋势，据此可以探索数据之间的关联关系，甚至选择合适的函数对数据点进行拟合。

2.2.2 相关数据来源



【例 2.2】为了解某高校新入学男生的身高及体重情况，某课题组随机抽取了该校新入学的 42 名大一新生的身高及体重数据，如表 2.2 所示。试通过绘制散点图来直观地反映这些学生的身高、体重组合情况。

表 2.2 某高校的 42 名大一新生的身高及体重

编号	身高/cm	体重/kg
1	176	67
2	185	77
3	177	77
4	165	59
5	174	64
…	…	…
40	173	66
41	172	63
42	174	60

2.2.3 Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有两个变量，分别是身高和体重。我们把身高变量设定为 SG，把体重变量设定为 TZ，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作我们在第 1 章中已有详细讲述。录入完成后，数据如图 2.7 所示。

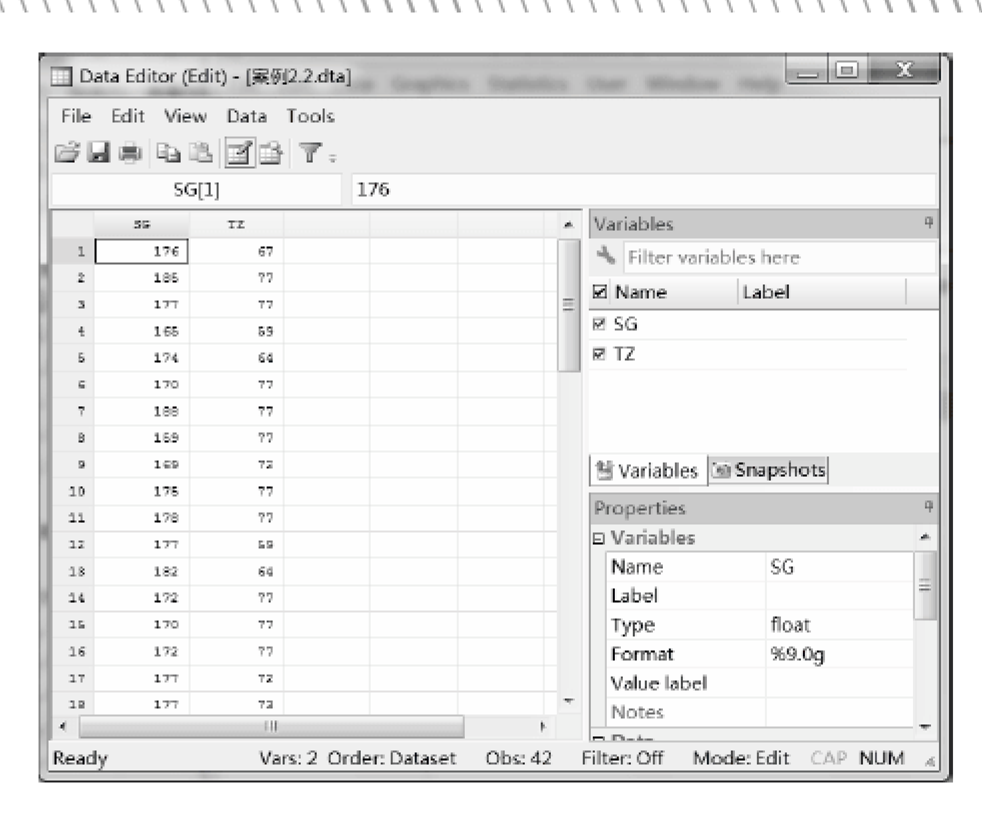


图 2.7　案例 2.2 数据

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关的数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入命令：

```
graph twoway scatter  SG TZ
```

03 设置完毕后，按键盘上的回车键，等待输出结果。

2.2.4　结果分析

上述操作结束后，Stata 14.0 将弹出如图 2.8 所示的散点图。

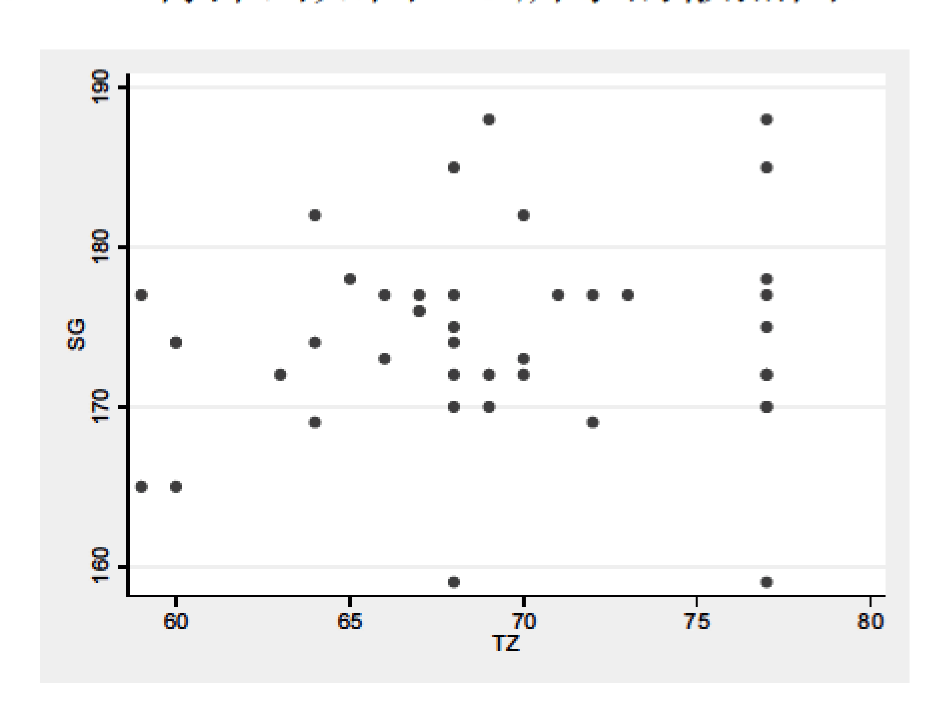


图 2.8　散点图 1

通过观察散点图，可以比较轻松地看出这些学生的身高及体重的组合情况。我们发现，大部分学生的身高处于 170cm~180cm 之间，身高与体重之间不存在明显的相关关系，很多体重差别较大的学生身高几乎无差别，同时有很多体重相近的学生之间身高差别很大。

2.2.5　案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

1．延伸 1：给图形增加标题、给坐标轴增加数值标签并设定间距、显示坐标轴的刻度

例如，我们要给图形增加标题的名称“案例 2.2 结果”，对 X 轴添加数值标签，取值为 56~80，间距为 2，对 Y 轴添加数值标签，取值为 150~190，间距为 10，对 Y 轴添加刻度，间距为 5，那么操作命令就应该相应地修改为：

```
graph twoway scatter  SG TZ,title("案例 2.2 结果")
xlabel(56(2)80) ylabel(150(10)190) ytick(150(5)190)
```

在命令窗口输入命令并按回车键进行确认，结果如图 2.9 所示。

2．延伸 2：控制散点标志的形状

例如，我们要在延伸 1 的基础上使散点图中散点标志的形状变为实心菱形，那么操作命令就应该相应地修改为：

```
graph twoway scatter  SG TZ,title("案例 2.2 结果")
xlabel(56(2)80) ylabel(150(10)190) ytick(150(5)190) msymbol(D)
```

在命令窗口输入命令并按回车键进行确认，结果如图 2.10 所示。

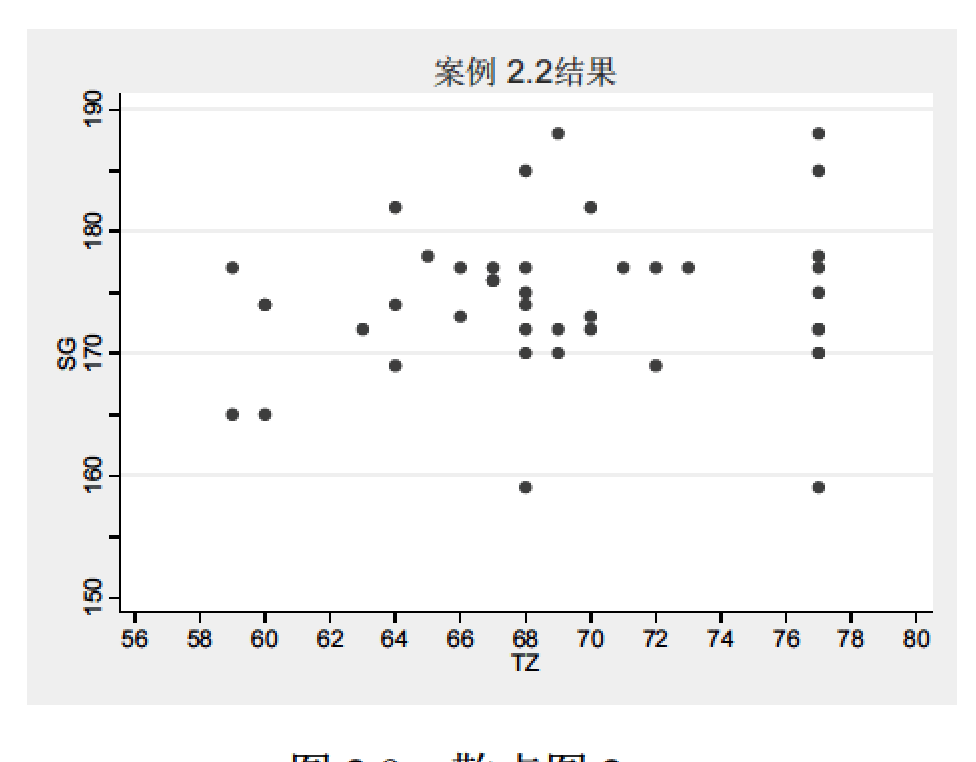



图 2.9　散点图 2

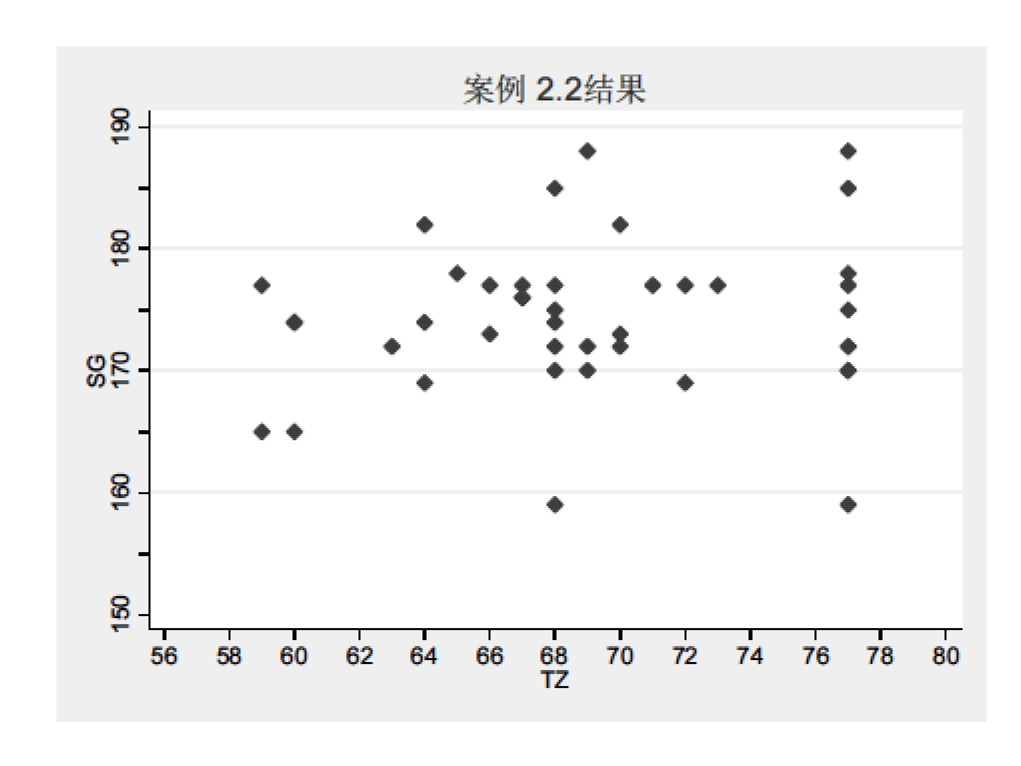


图 2.10　散点图 3

在上面的例子中，命令中的 D 代表的是实心菱形。散点标志的其他常用可选形状与对应命令缩写如表 2.3 所示。

表 2.3　形状与对应命令

缩写	描述	缩写	描述	缩写	描述
X	大写字母X	S	实心方形	th	空心小三角形
Th	空心三角	oh	空心小圆圈	sh	空心方形
T	实心三角	p	很小的点	dh	空心小菱形

3．延伸 3：控制散点标志的颜色

例如，我们要在延伸 2 的基础上进行改进，使散点标志的颜色变为黄色，那么操作命令就应该相应地修改为：

```
graph twoway scatter  SG TZ,title("案例 2.2 结果")
xlabel(56(2)80) ylabel(150(10)190) ytick(150(5)190) msymbol(D) mcolor(yellow)
```

在命令窗口输入命令并按回车键进行确认，结果如图 2.11 所示。

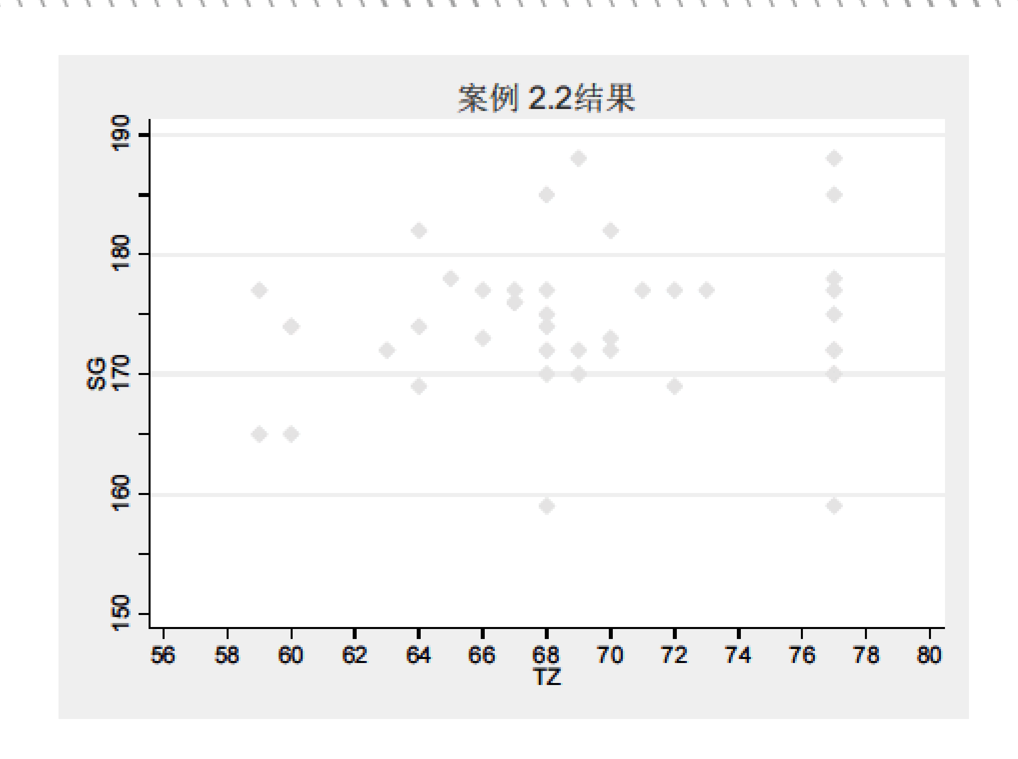


图 2.11　散点图 4

更多颜色选择，请在命令窗口输入命令：

```
help colorstyle
```

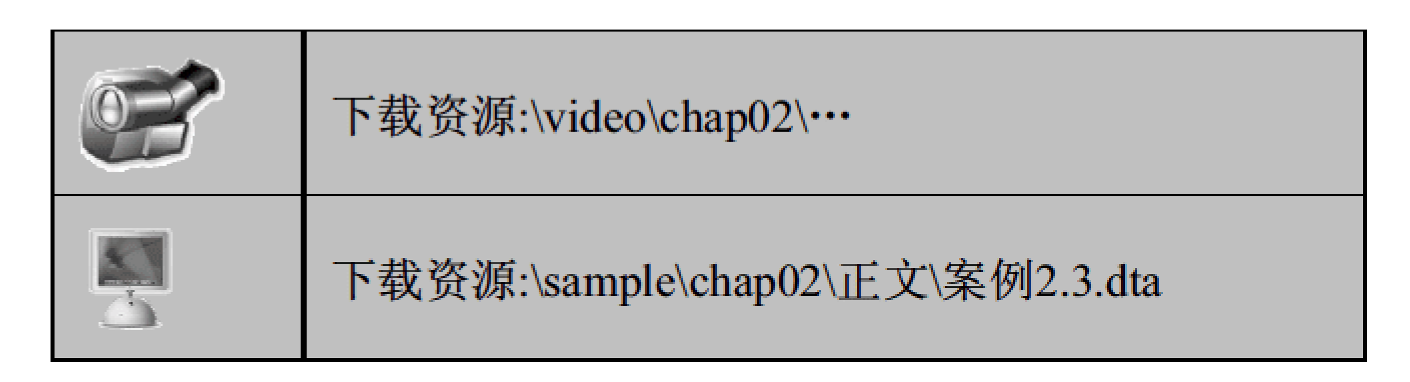
然后按回车键进行确认即可选择。

2.3　实例三——曲线标绘图

2.3.1　曲线标绘图的功能与意义

从形式上看，曲线标绘图与散点图的区别就是一条线来替代散点标志，这样做可以更加清晰直观地看出数据走势，但却无法观察到每个散点的准确定位。从用途上看，曲线标绘图常用于时间序列分析的数据预处理，用来观察变量随时间的变化趋势。此外，曲线标绘图可以同时反映多个变量随时间的变化情况，所以，曲线标绘图的应用范围还是非常广泛的。

2.3.2　相关数据来源


下载资源:\video\chap02\…

下载资源:\sample\chap02\正文\案例2.3.dta

【例 2.3】某足球教练准备执教一支新球队，在执教前对拟执教球队的过往赛季进球数据进行了搜集整理，如表 2.4 所示。试通过绘制曲线标绘图来分析研究该球队的进球情况变化趋势以及对队内第 1 射手（进球最多的队员）的依赖度。

表 2.4　拟执教球队的过往赛季进球数据

年份	球队总进球数	球队第1射手进球数
1997	69	15
1998	68	16
1999	74	16

（续表）

年份	球队总进球数	球队第1射手进球数
2000	73	17
2001	59	21
…	…	…
2010	68	39
2011	70	38
2012	71	41

2.3.3 Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有 3 个变量，分别是年份、总进球数和第 1 射手进球数。我们把年份变量设定为 year，把总进球数变量设定为 total，把第 1 射手进球数变量设定为 first，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作我们在第 1 章中已有详细讲述。录入完成后数据如图 2.12 所示。

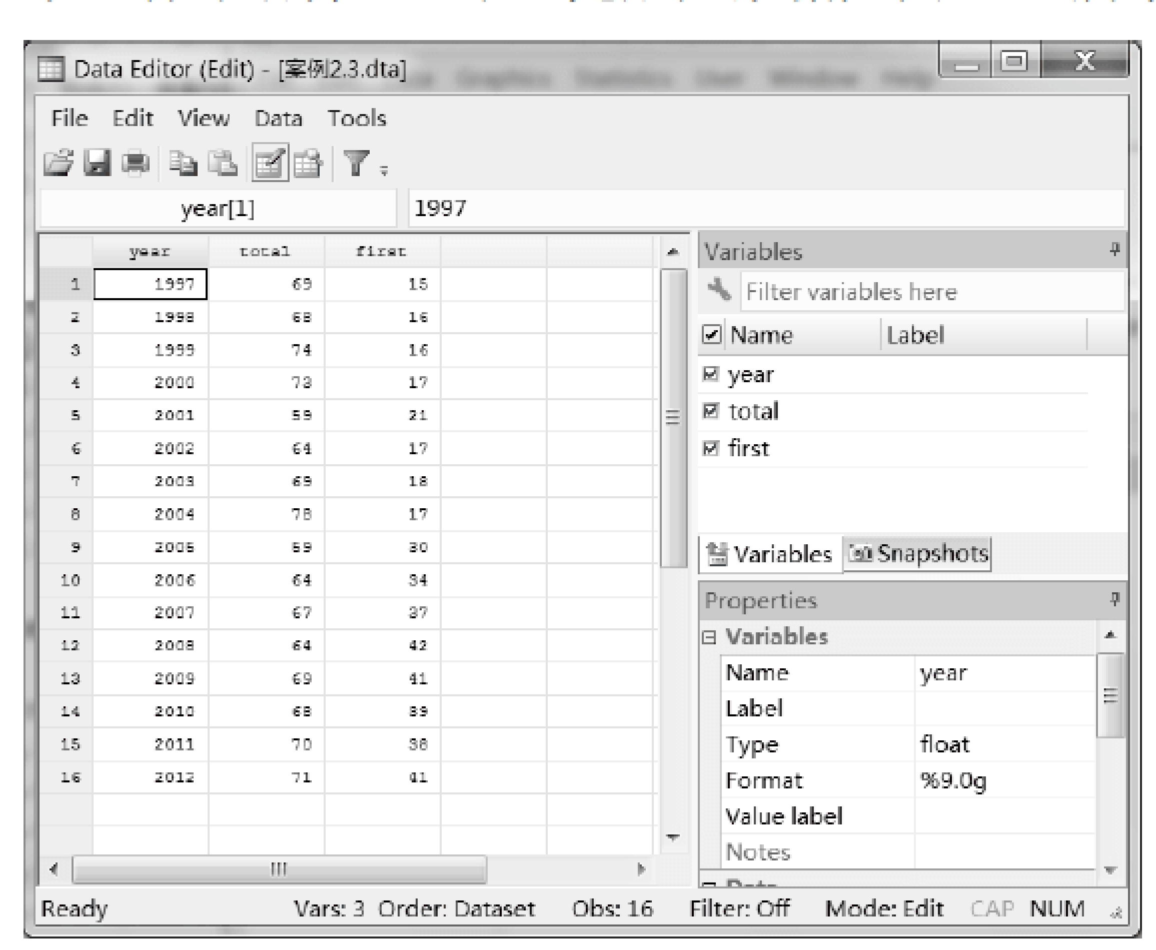


图 2.12　案例 2.3 数据

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入命令：

```
graph twoway line  total first year
```

03 设置完毕后，按键盘上的回车键，等待输出结果。

2.3.4 结果分析

上述操作完成后，Stata 14.0 将弹出如图 2.13 所示的曲线标绘图。

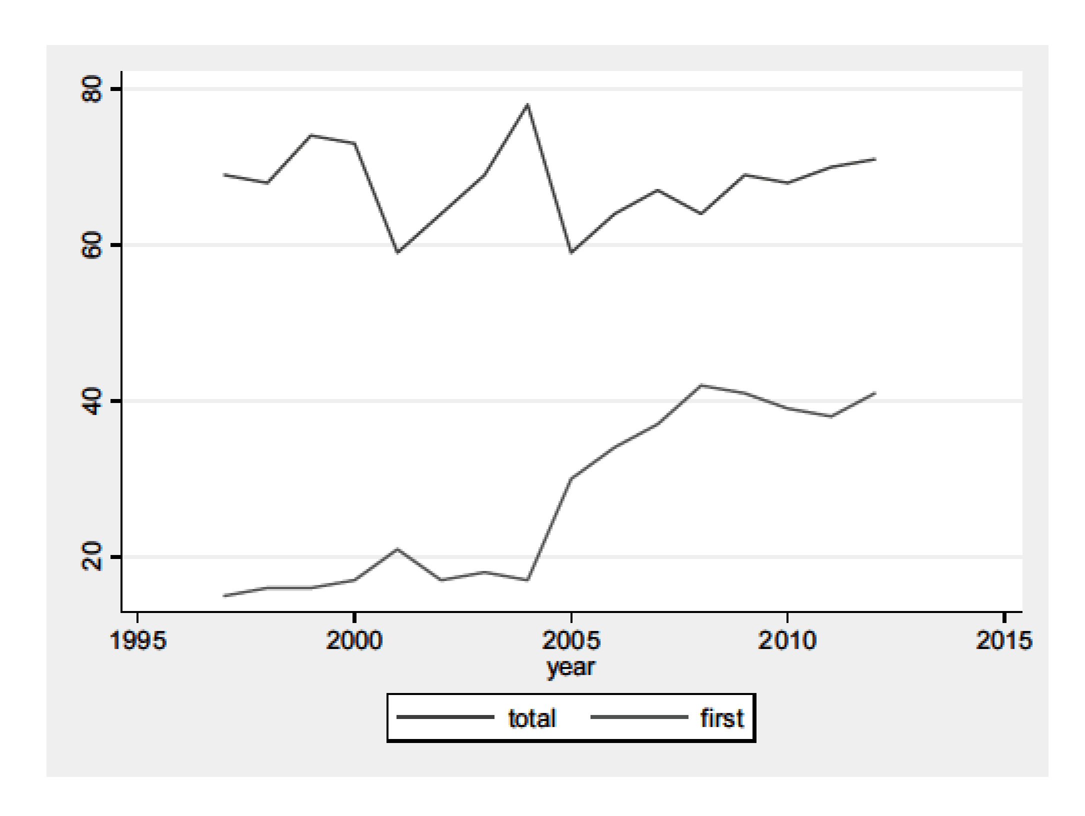


图 2.13　曲线标绘图 1

通过观察曲线图，可以比较轻松地看出本支球队的进球情况。我们发现，该球队的进球状态虽有所起伏却变化不大，但是队内第 1 射手的进球状态是在波动中上升的，这可能是原来的射手逐渐成熟、成长起来，能力得到提升，也有可能是引进了更加优秀的球员所致。从整体上看，该支球队并没有完全依赖第 1 射手进球，但是它的依赖度自 2005 年以来是有所上升的。

2.3.5　案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

1．延伸 1：给图形增加标题、给坐标轴增加数值标签并设定间距、显示坐标轴的刻度

例如我们要给图形增加标题的名称“案例 2.3 结果”，对 X 轴添加数值标签，取值为 1997~2012，间距为 2，对 Y 轴添加数值标签，取值为 0~80，间距为 10，对 X 轴添加刻度，间距为 1，那么操作命令就应该相应地修改为：

```
graph twoway line  total first year,title("案例 2.3 结果") xlabel(1997(2)2012)
ylabel(0(10)80) xtick(1997(1)2012)
```

在命令窗口输入命令并按回车键进行确认，结果如图 2.14 所示。

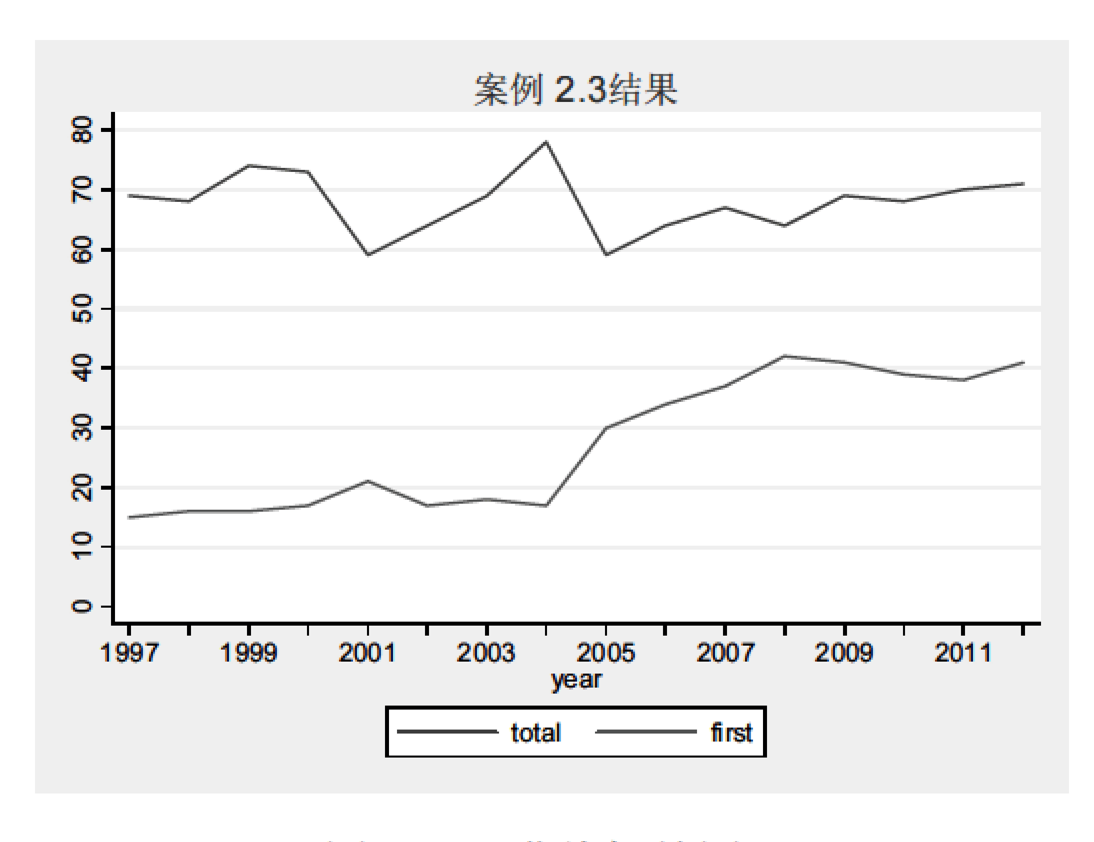


图 2.14　曲线标绘图 2

2．延伸 2：改变变量默认标签

例如，我们要在延伸 1 的基础上使总进球数和第 1 射手进球数这两个变量的标签直接以汉字显示，从而更加清晰直观，那么操作命令就应该相应地修改为：

```
graph twoway line  total first year,title("案例 2.3 结果") xlabel(1997(2)2012)
ylabel(0(10)80) xtick(1997(1)2012) legend(label(1 "总进球数")
label(2 "第 1 射手进球数"))
```

在命令窗口输入命令并按回车键进行确认，结果如图 2.15 所示。

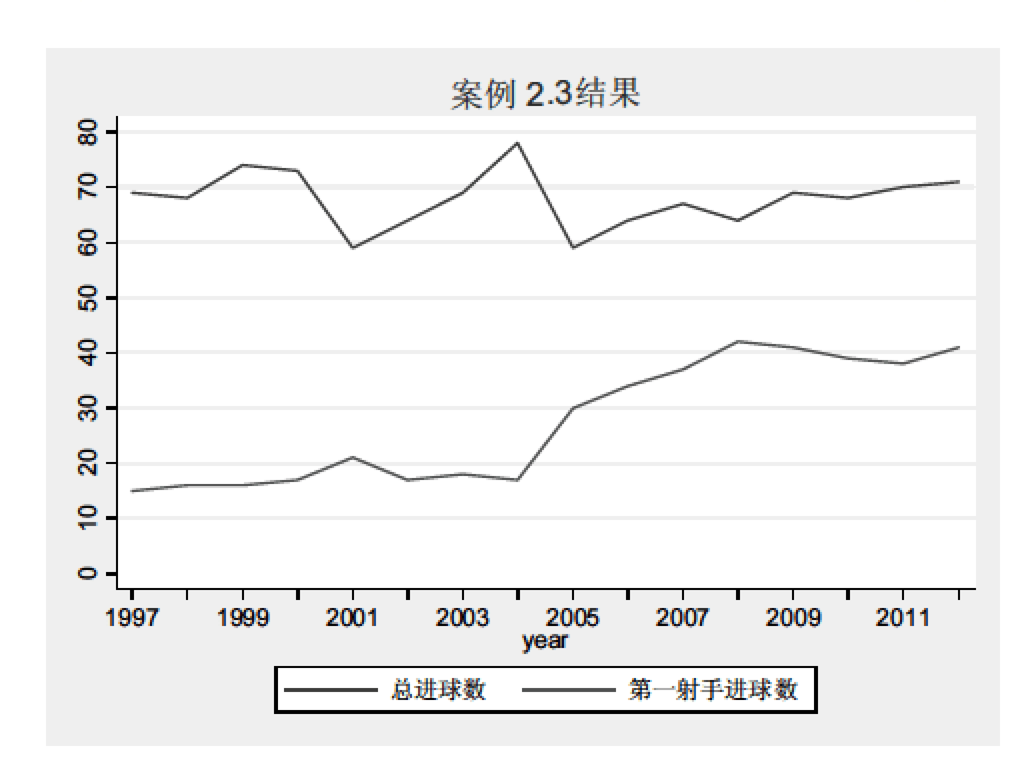


图 2.15　曲线标绘图 3

3．延伸 3：改变线条的样式

例如，我们要在延伸 2 的基础上进行改进，使第 1 射手进球数的曲线变为虚线，那么操作命令就应该相应地修改为：

```
graph twoway line  total first year,title("案例 2.3 结果") xlabel(1997(2)2012)
ylabel(0(10)80) xtick(1997(1)2012) legend(label(1 "总进球数")
label(2 "第 1 射手进球数")) clpattern(solid dash)
```

在命令窗口输入命令并按回车键进行确认，结果如图 2.16 所示。

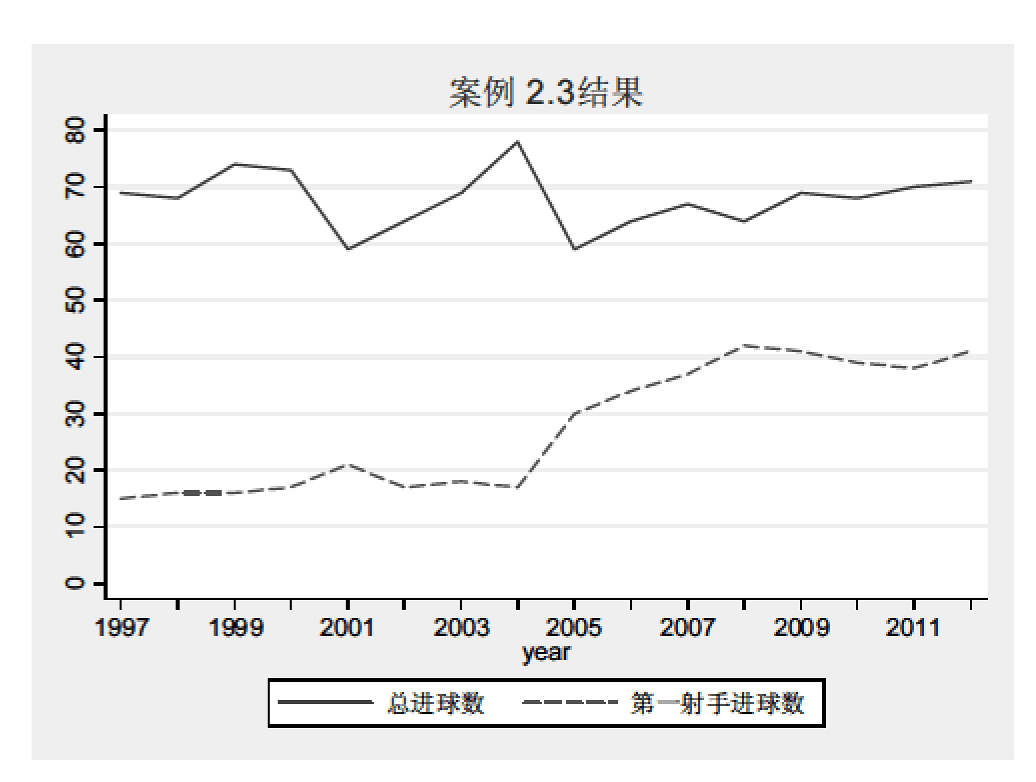


图 2.16　曲线标绘图 4

在上述命令中，solid 代表实线，对应的是第 1 个因变量 total；dash 代表虚线，对应的是第 2 个因变量 first。线条样式与其对应的命令缩写如表 2.5 所示。

表 2.5　线条样式与命令缩写

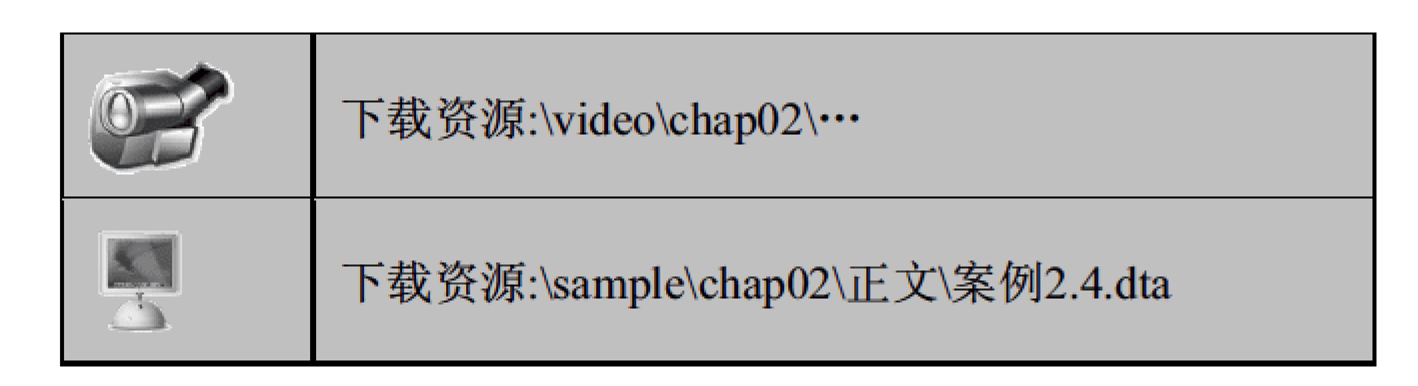
线条样式	命令缩写	线条样式	命令缩写	线条样式	命令缩写
实线	solid	点划线	dash_dot	长划线	longdash
虚线	dash	短划线	shortdash	长划点线	longdash_dot
点线	line	短划点线	shortdash_dot	不可见的线	blank

2.4　实例四——连线标绘图

2.4.1　连线标绘图的功能与意义

在 2.3 节中我们提到曲线标绘图用一条线来代替散点标志，可以更加清晰直观地看出数据走势，但却无法观察到每个散点的准确定位。那么，有没有一种作图方式既可以满足观测数据走势的需要，又能实现每个散点的准确定位？Stata 的连线标绘图制图方法就提供了解决这一问题的方法。

2.4.2　相关数据来源


下载资源:\video\chap02\…

下载资源:\sample\chap02\正文\案例2.4.dta

【例 2.4】A 市旅游局决定对辖区内某一王牌旅游景点进行游客量调查，调查得到的数据经整理后如表 2.6 所示。试通过绘制连线标绘图来分析研究该景点的游客量随季节的变化情况。

表 2.6　某旅游景点各月份旅游人次

月份	游客量/人/次
1	1779
2	2339
3	2559
4	3429
5	5689
…	…
10	6798
11	2794
12	1986

2.4.3　Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有两个变量，分别是月份、游客量。我们把月份变量设定为 month，把游客量变量设定为 number，变量类型及长度

采取系统默认方式，然后录入相关数据。相关操作我们在第 1 章中已有详细讲述。录入完成后数据如图 2.17 所示。

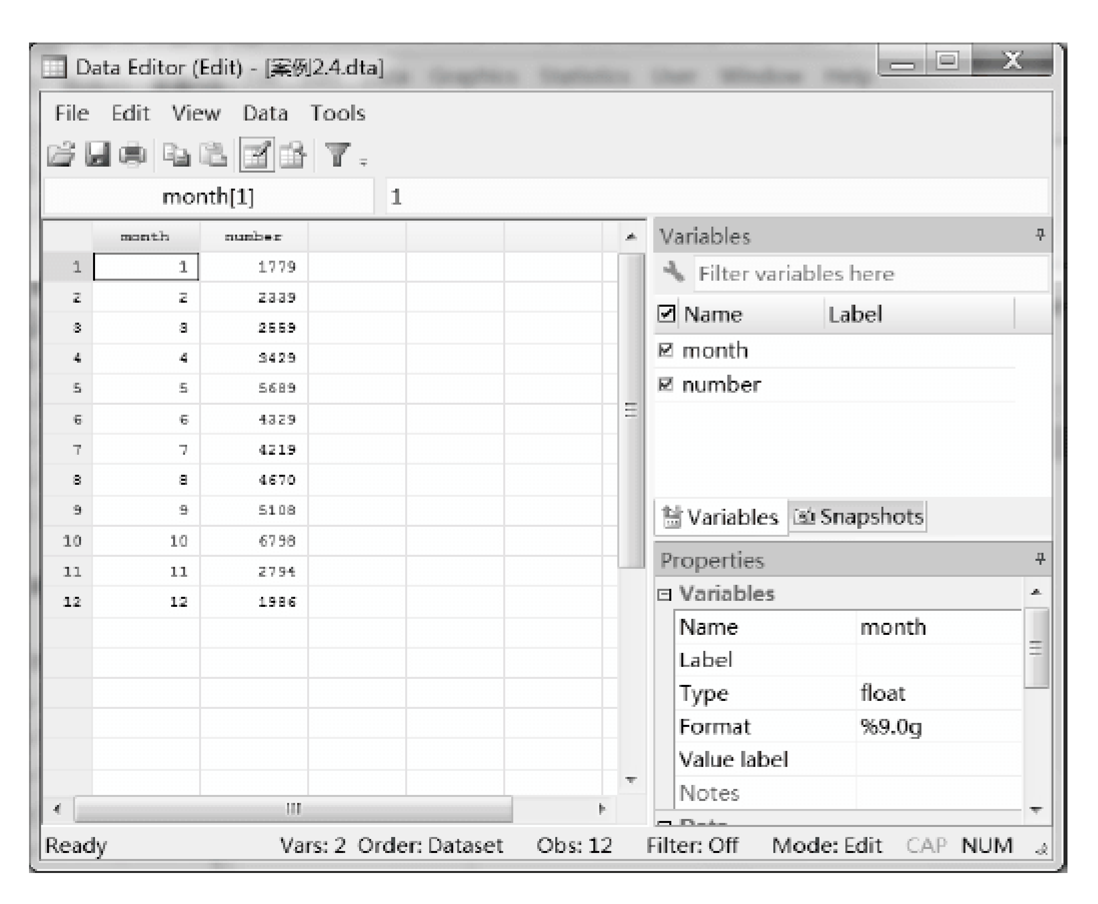


图 2.17　案例 2.4 数据

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入命令：

```
graph twoway connected  number month
```

03 设置完毕后，按键盘上的回车键，等待输出结果。

2.4.4　结果分析

上述操作完成后，Stata 14.0 将弹出如图 2.18 所示的连线标绘图。

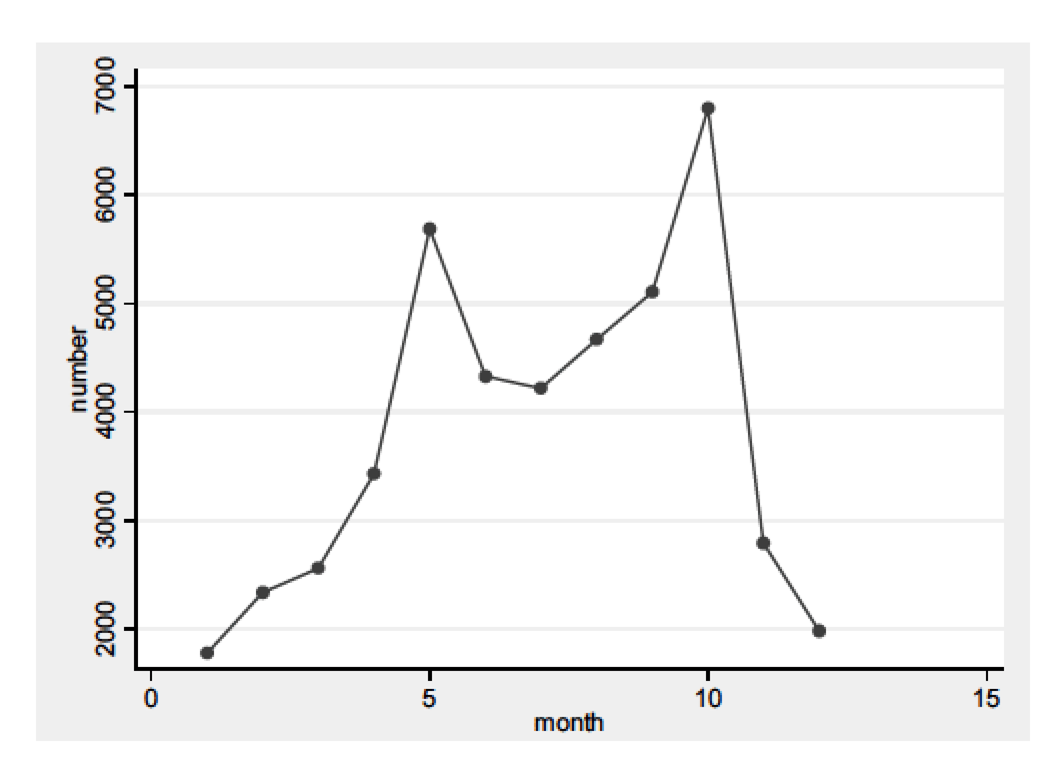


图 2.18　连线标绘图 1

通过观察连线标绘图，我们可以了解到很多信息：一方面可以清晰地看到该景点各个月份的游客人次的准确值；另一方面可以看到该景点游客人次的变化趋势。该景点的 5~10 月份是旺季，其中 10 月份游客人数最多，其他的月份属于淡季，1 月份的游客人数最低。决策者可以根据这一规律为景点合理配置资源、制定差别价格等。

2.4.5 案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

1．延伸 1：给图形增加标题、给坐标轴增加数值标签并设定间距、显示坐标轴的刻度

例如，我们要给图形增加标题的名称“案例 2.4 结果”，对 X 轴添加数值标签，取值为 1~12，间距为 1，对 Y 轴添加数值标签，取值为 1000~7000，间距为 1000，对 Y 轴添加刻度，间距为 500，那么操作命令就应该相应地修改为：

```
graph twoway connected  number month,title("案例 2.4 结果") xlabel(1(1)12)
ylabel(1000(1000)7000) ytick(1000(500)7000)
```

在命令窗口输入命令并按回车键进行确认，结果如图 2.19 所示。

2．延伸 2：改变线条的样式

例如，我们要在延伸 1 的基础上进行改进，使游客量的曲线变为虚线，那么操作命令就应该相应地修改为：

```
graph twoway connected  number month,title("案例 2.4 结果") xlabel(1(1)12)
ylabel(1000(1000)7000) ytick(1000(500)7000) clpattern(dash)
```

在命令窗口输入命令并按回车键进行确认，结果如图 2.20 所示。

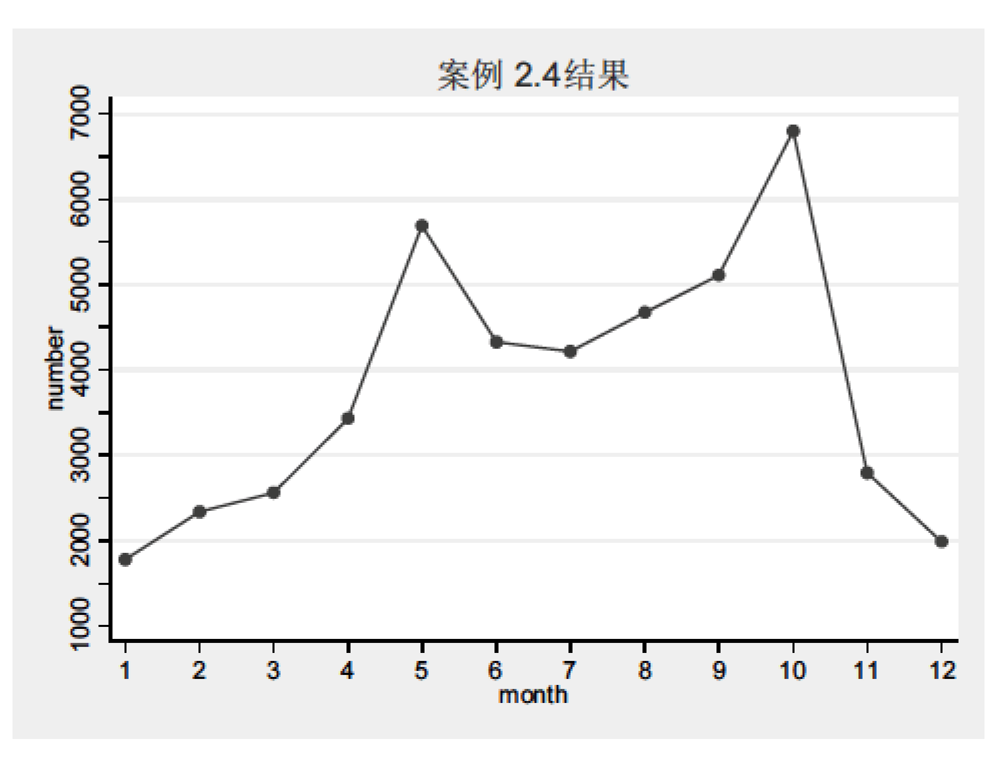


图 2.19 连线标绘图 2

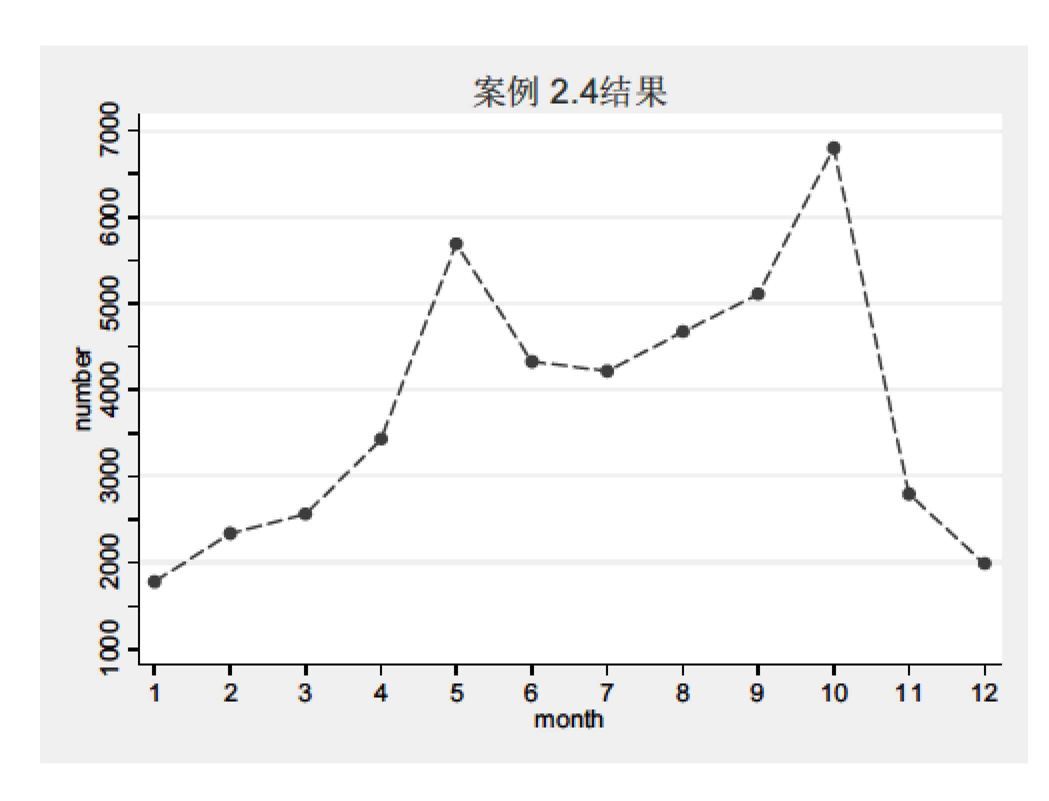


图 2.20 连线标绘图 3

3．延伸 3：控制散点标志的形状

例如，我们要在延伸 2 的基础上使连线标绘图中散点标志的形状变为实心菱形，那么操作命令就应该相应地修改为：

```
graph twoway connected  number month,title("案例 2.4 结果") xlabel(1(1)12)
ylabel(1000(1000)7000) ytick(1000(500)7000) clpattern(dash) msymbol(D)
```

在命令窗口输入命令并按回车键进行确认，结果如图 2.21 所示。

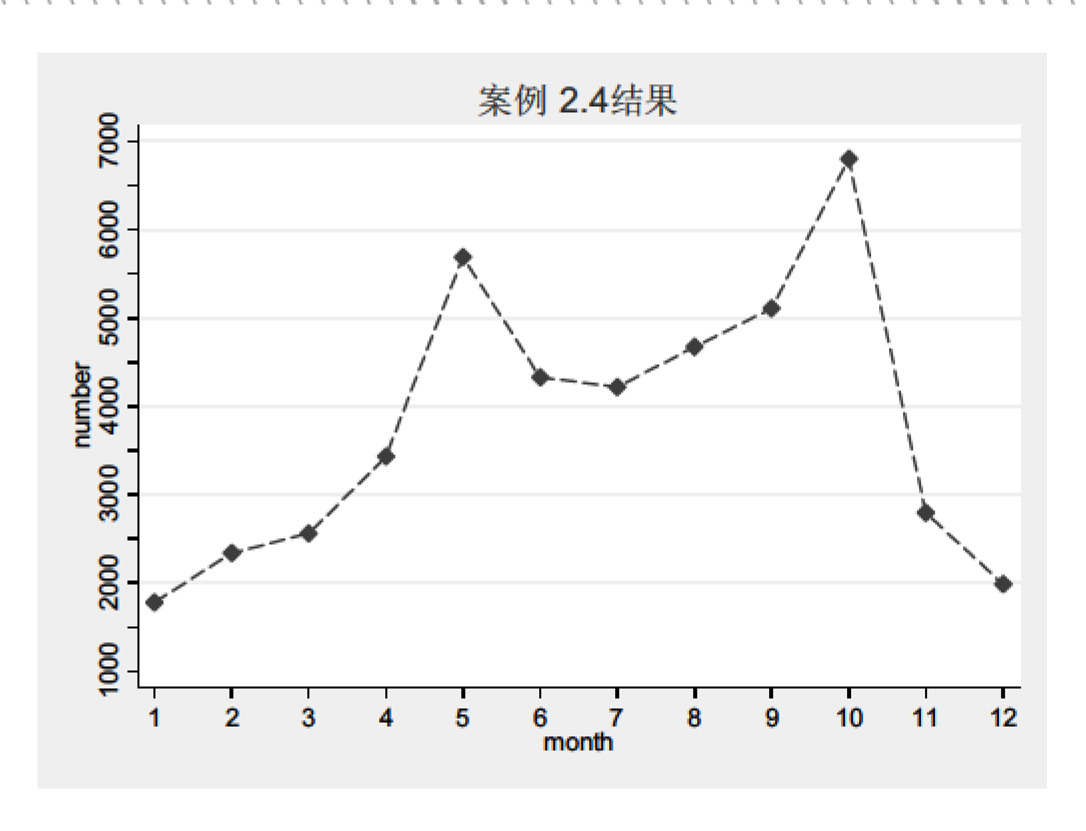


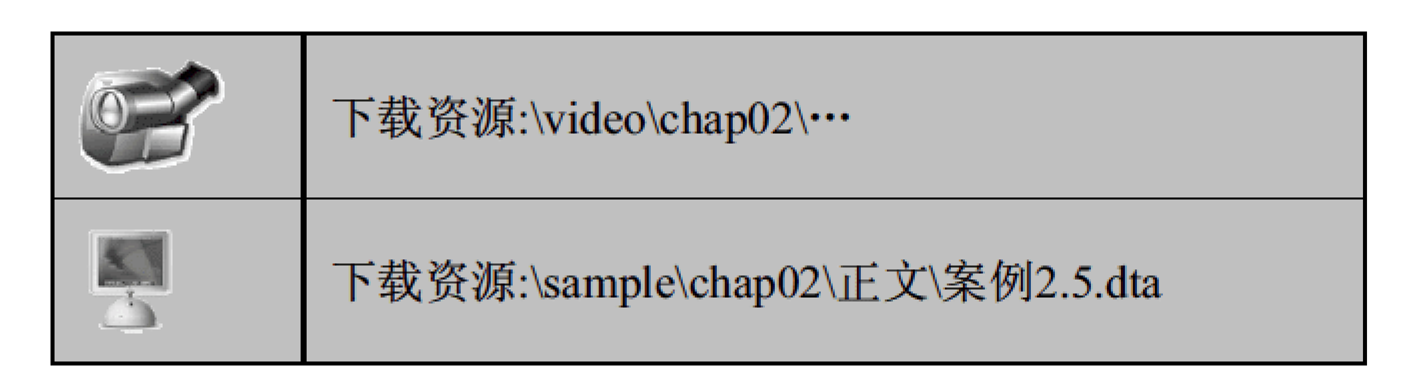
图 2.21　连线标绘图 4

2.5　实例五——箱图

2.5.1　箱图的功能与意义

箱图（Box-Plot）又称为盒须图、盒式图或箱线图，是一种用于显示一组数据分散情况的统计图。箱图很形象地分为中心、延伸以及分部状态的全部范围，提供了一种只用 5 个点对数据集做简单总结的方式，这 5 个点包括中点、Q1、Q3、分部状态的高位和低位。数据分析者通过绘制箱图不仅可以直观明了地识别数据中的异常值，判断数据的偏态、尾重以及比较几批数据的形状。

2.5.2　相关数据来源



	下载资源:\video\chap02\···
	下载资源:\sample\chap02\正文\案例2.5.dta

【例 2.5】X 集团是一家国内大型销售汽车的公司，该公司在组织架构上采取的是事业部制管理方式，把全国市场分为 3 个大区，从而督导各省市的分公司。该集团在全国各省市的市场份额情况如表 2.7 所示。试绘制箱图来研究分析其分布规律。

表 2.7　某集团各大分区的市场份额情况

地区	市场份额	所属大区
北京	38	1
天津	44	1
河北	22	1
山西	8	1
内蒙古	32	1
…	…	…

（续表）

地区	市场份额	所属大区
青海	18	3
宁夏	20	3
新疆	60	3

2.5.3　Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有 3 个变量，分别是地区、市场份额以及所属大区。我们把地区变量设定为 region，把市场份额设定为 SCFE，把所属大区变量设定为 Center，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作我们在第 1 章中已有详细讲述。录入完成后数据如图 2.22 所示。

先做一下数据保存，然后开始展开分析步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入命令：

```
graph box SCFE
```

03 设置完毕后，按键盘上的回车键，等待输出结果。

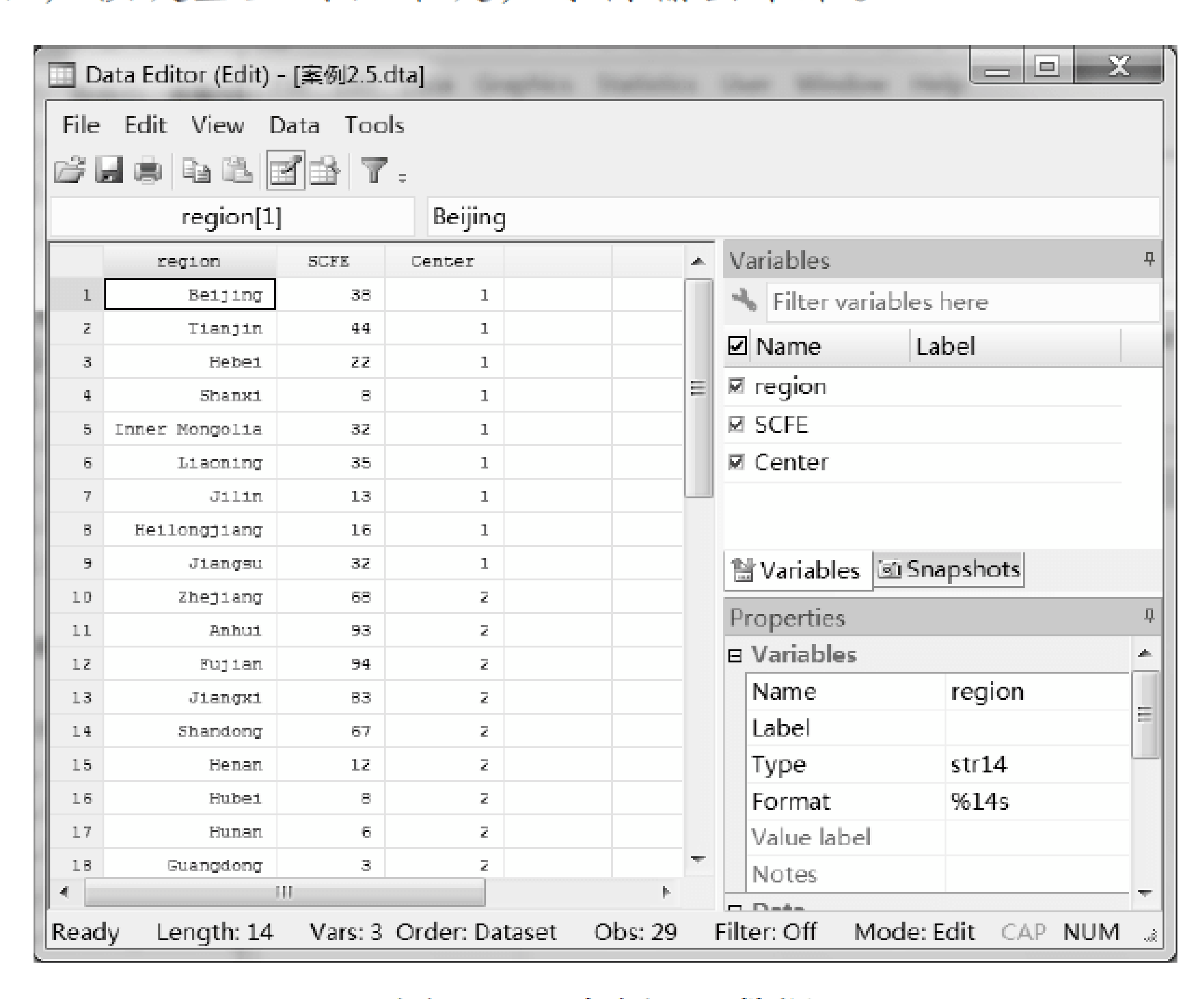


图 2.22　案例 2.5 数据

2.5.4　结果分析

上述操作完成后，Stata 14.0 将弹出如图 2.23 所示的箱图。

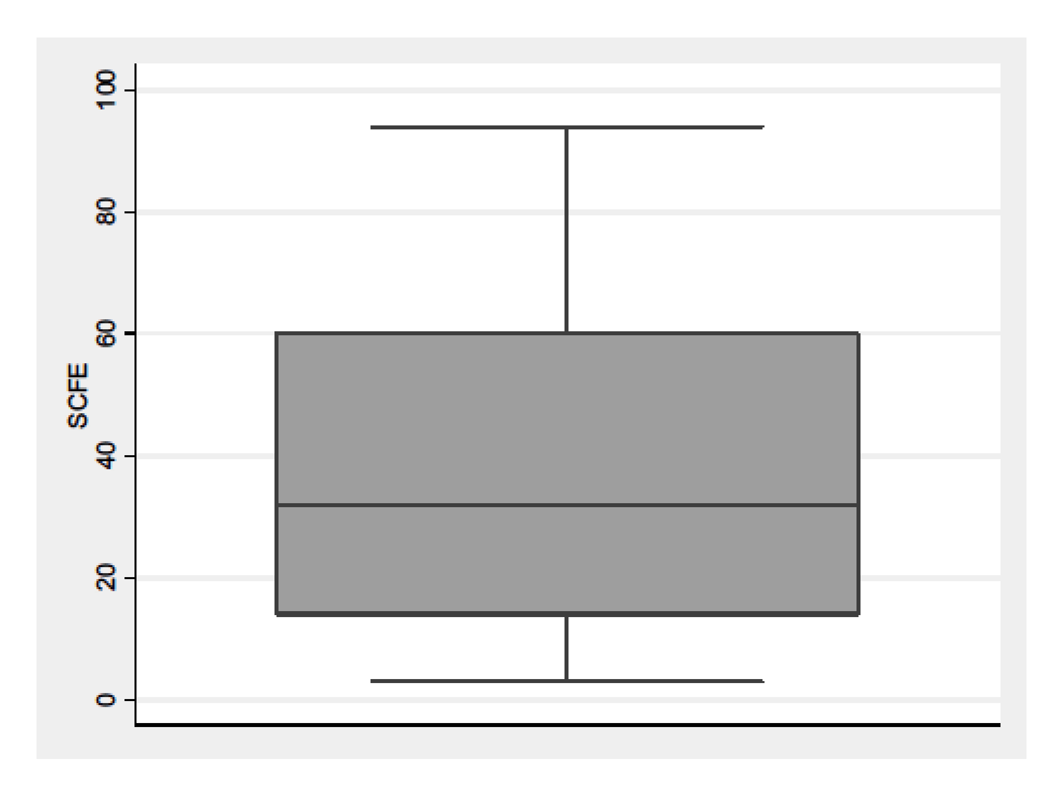


图 2.23　箱图 1

通过观察箱图，可以了解到很多信息。箱图把所有的数据分成了 4 部分，第 1 部分是从顶线到箱子的上部，这部分数据值在全体数据中排名前 25%；第 2 部分是从箱子的上部到箱子中间的线，这部分数据值在全体数据中排名 25%以下，50%以上；第 3 部分是从箱子中间的线到箱子的下部，这部分数据值在全体数据中排名 50%以下，75%以上；第 4 部分是从箱子的底部到底线，这部分数据值在全体数据中排名后 25%。顶线与底线的间距在一定程度上表示了数据的离散程度，间距越大就越离散。就本例而言，我们可以看到该公司市场份额的中位数在 32%左右，市场份额最高的省市可达到 90%左右。

2.5.5　案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

延伸：我们能否把上面各省市的市场份额数据按照所属各个大区分别绘制箱图呢？答案是肯定的。

操作命令应该相应地修改为：

```
graph box  SCFE,over( Center)
```

在命令窗口输入命令并按回车键进行确认，结果如图 2.24 所示。

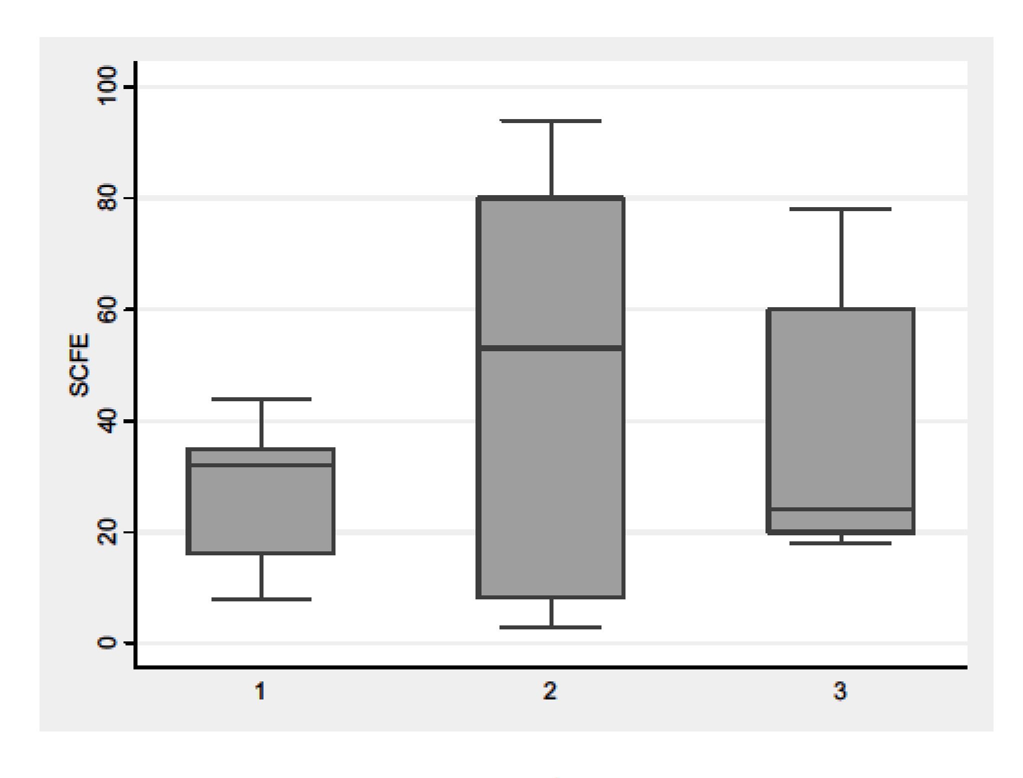


图 2.24　箱图 2

从该图中可以看出第 2 大区的市场份额中位数水平是最高的，第 3 大区的市场份额中位数水平最低，第 1 大区的市场份额中位数水平居中。第 2 大区各个省市之间的市场份额情况相对存在较大差异。

2.6　实例六——饼图

2.6.1　饼图的功能与意义

饼图是数据分析中常见的一种经典图形，因其外形类似于圆饼而得名。在数据分析中，很多时候需要分析数据总体的各个组成部分的占比，我们可以通过各个部分与总额相除来计算，但这种数学比例的表示方法相对抽象，Stata 14.0 提供了饼形制图工具，能够直接以图形的方式显示各个组成部分所占比例，更为重要的是，由于采用图形的方式，因此更加形象直观。

2.6.2　相关数据来源

	下载资源:\video\chap02\···
	下载资源:\sample\chap02\正文\案例2.6.dta

【例 2.6】B 股份有限公司是一家资产规模巨大的国内上市公司，公司采取多元化经营的成长型发展战略，经营范围包括餐饮、房地产、制造等，公司采取区域事业部制的组织架构，在东部、中部、西部都有自己的分部，较为独立地负责本部各产业的具体运营。该公司各大分部的具体营业收入数据如表 2.8 所示。试通过绘制饼图的方式研究该公司各产业的占比情况。

表 2.8　某集团各大分部的市场份额情况

地区	餐饮业营业收入/万元	房地产业营业收入/万元	制造业营业收入/万元
东部	2089	9845	10234
中部	828	6432	7712
西部	341	1098	1063

2.6.3　Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有 4 个变量，分别是地区、餐饮业营业收入、房地产业营业收入以及制造业营业收入。我们把地区变量设定为 region，把餐饮业营业收入变量设定为 CANYIN，把房地产业营业收入变量设定为 FANGCHAN，把制造业营业收入变量设定为 ZHIZAO，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作我们在第 1 章中已有详细讲述。录入完成后数据如图 2.25 所示。

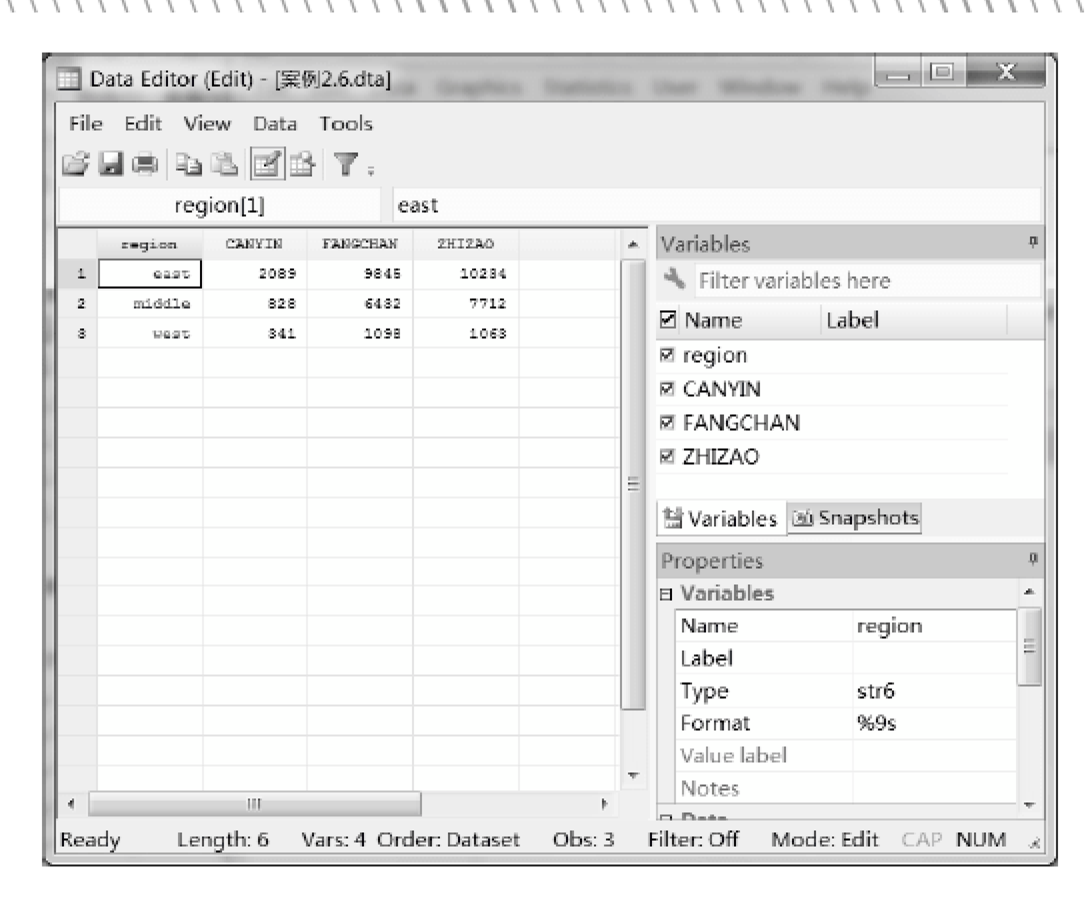


图 2.25　案例 2.6 数据

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入命令：

```
graph pie CANYIN FANGCHAN ZHIZAO
```

03 设置完毕后，按键盘上的回车键，等待输出结果。

2.6.4　结果分析

上述操作完成后，Stata 14.0 会弹出如图 2.26 所示的饼图。

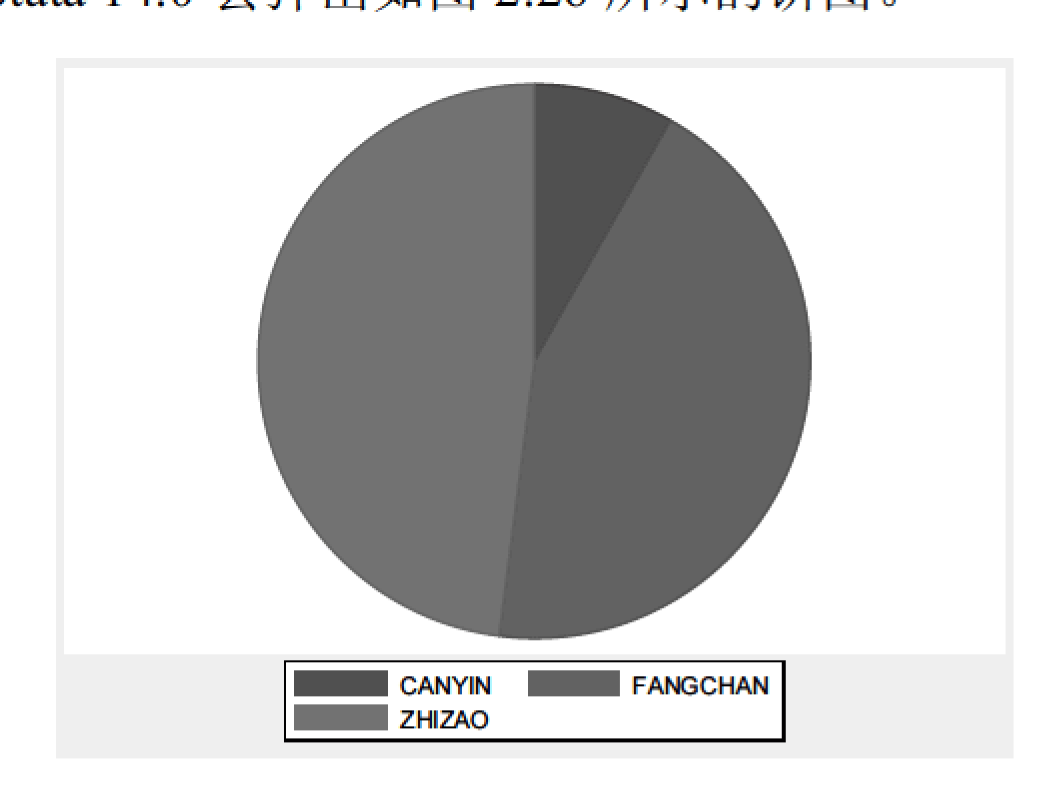


图 2.26　饼图 1

通过观察饼图，我们可以比较轻松地看出企业的主营业务，该企业的两个支柱产业是制造业和房地产，餐饮业占比较小。

2.6.5　案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

1．延伸 1：对图形展示进行更加个性化的设置

例如，我们要把餐饮业的营业收入占比突出显示，把房地产业营业收入的饼颜色改为黄色，给餐饮业营业收入和房地产业营业收入的饼在距中心 20 个相对半径单位的位置处加上百分比标签，那么操作命令就应该相应地修改为：

```
graph pie  CANYIN FANGCHAN ZHIZAO,pie(1,explode) pie(2,color(yellow))
plabel(1 percent,gap(20)) plabel(2 percent,gap(20))
```

在命令窗口输入命令并按回车键进行确认，结果如图 2.27 所示。

2．延伸 2：按照分类变量分别画出饼图

例如，我们要在延伸 1 的基础上通过绘制饼图的方式研究该公司每个分部内各个产业的占比情况，那么操作命令就应该相应地修改为：

```
graph pie  CANYIN FANGCHAN ZHIZAO,pie(1,explode) pie(2,color(yellow))
plabel(1 percent,gap(20)) plabel(2 percent,gap(20)) by( region)
```

在命令窗口输入命令并按回车键进行确认，结果如图 2.28 所示。

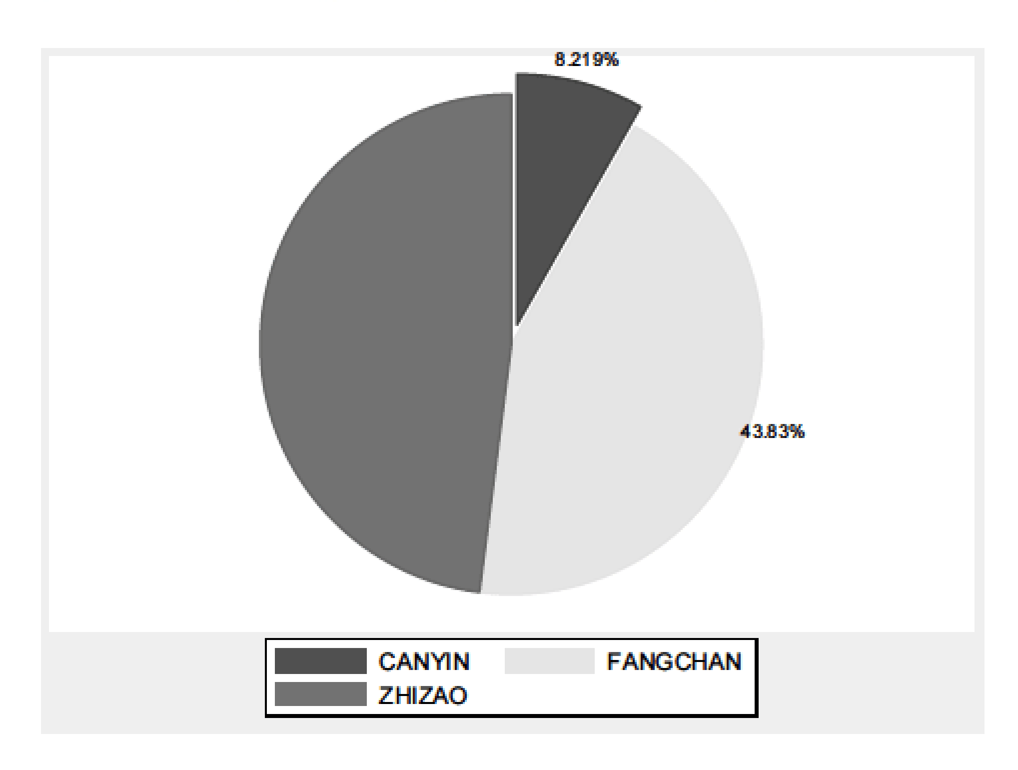


图 2.27　饼图 2

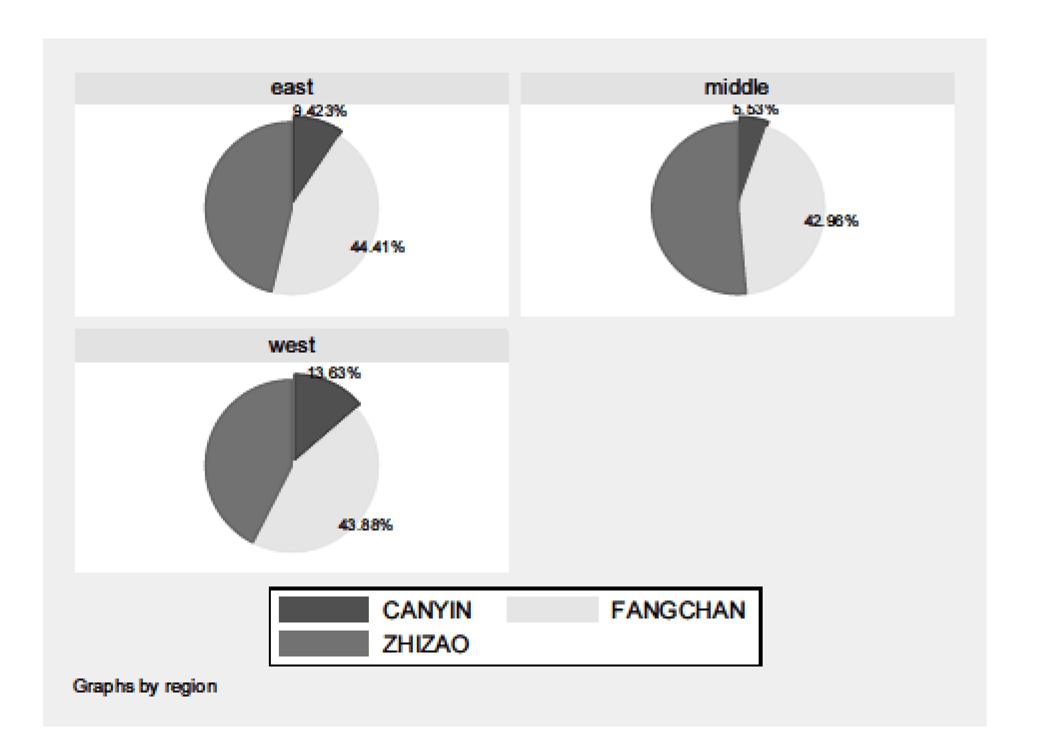


图 2.28　饼图 3

在上面的结果中，可以看到该公司每个分部各个产业的占比情况。例如，东部地区，观测左上方的 east 图就可以得到想要的答案。

2.7　实例七——条形图

2.7.1　条形图的功能与意义

相对于前面提到的箱图，条形图（Bar Chart）本身所包含的信息相对较少，但是它们仍然为平均数、中位数、合计数或计数等多种概要统计提供了简单又多样化的展示，所以条形图也深受研究者的喜爱，经常出现在研究者的论文或者调查报告中。

2.7.2 相关数据来源

	下载资源:\video\chap02\···
	下载资源:\sample\chap02\正文\案例2.7.dta

【例 2.7】某地方商业银行内设立 4 个营销团队，分别为 A、B、C、D，其营业净收入以及团队人数的具体情况如表2.9所示。试通过绘制条形图的方式来对比分析各团队的工作业绩。

表 2.9　某银行各营销团队营业净收入情况

营销团队	营业净收入/万元	团队人数/人
A	1899	1000
B	2359	1100
C	3490	1200
D	6824	1200

2.7.3 Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有 3 个变量，分别是团队名称、营业净收入以及团队人数。我们把团队名称变量设定为 team，把营业净收入变量设定为 sum，把团队人数变量设定为 number，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作我们在第 1 章中已有详细讲述。录入完成后数据如图 2.29 所示。

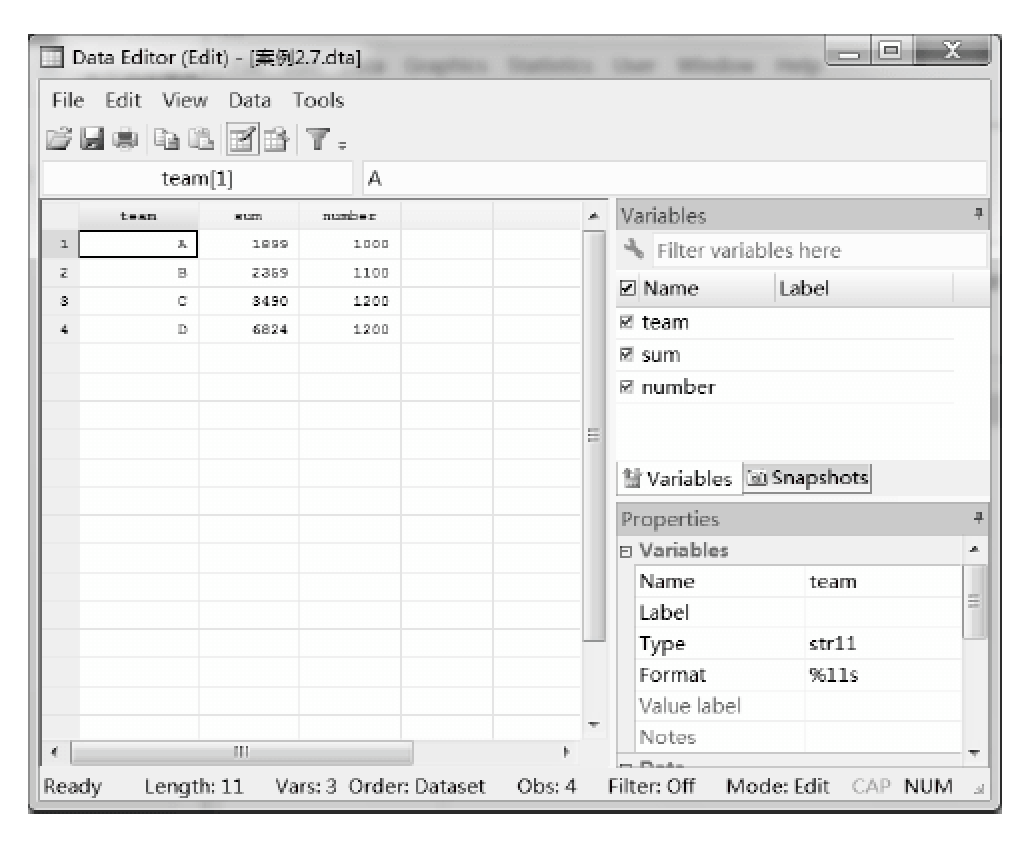


图 2.29　案例 2.7 数据

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入命令：

```
graph bar  sum,over( team)
```

03 设置完毕后，按键盘上的回车键，等待输出结果。

2.7.4　结果分析

上述操作完成后，Stata 14.0 会弹出如图 2.30 所示的条形图。

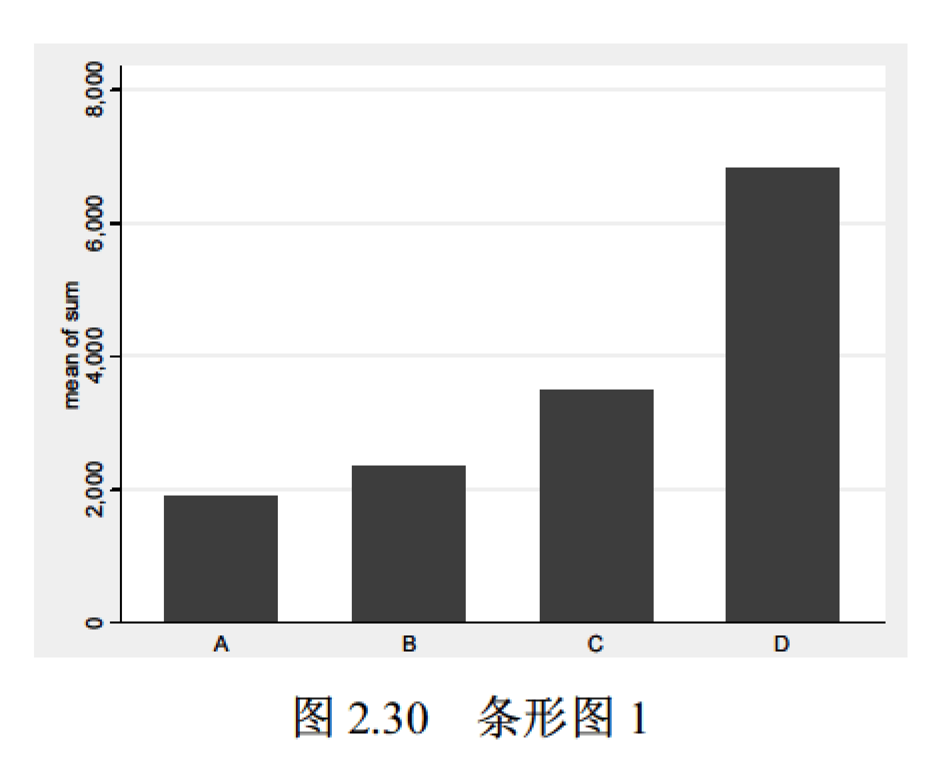


图 2.30　条形图 1

通过观察该条形图，我们可以比较轻松地看出该地方商业银行的 4 个团队的总体工作业绩，其中 D 团队成绩最好，C 其次，B 再次，A 最差。

2.7.5　案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

1．延伸 1：给图形增加标题、给坐标轴增加数值标签并设定间距、显示坐标轴的刻度

例如，我们要给图形增加标题的名称"案例 2.7 结果"，对 Y 轴添加数值标签，取值为 1000~7000，间距为 1000，对 Y 轴添加刻度，间距为 500，那么操作命令就应该相应地修改为：

```
graph bar sum,over(team) title("案例 2.7 结果") ylabel(1000(1000)7000)
ytick(1000(500)7000)
```

在命令窗口输入命令并按回车键进行确认，结果如图 2.31 所示。

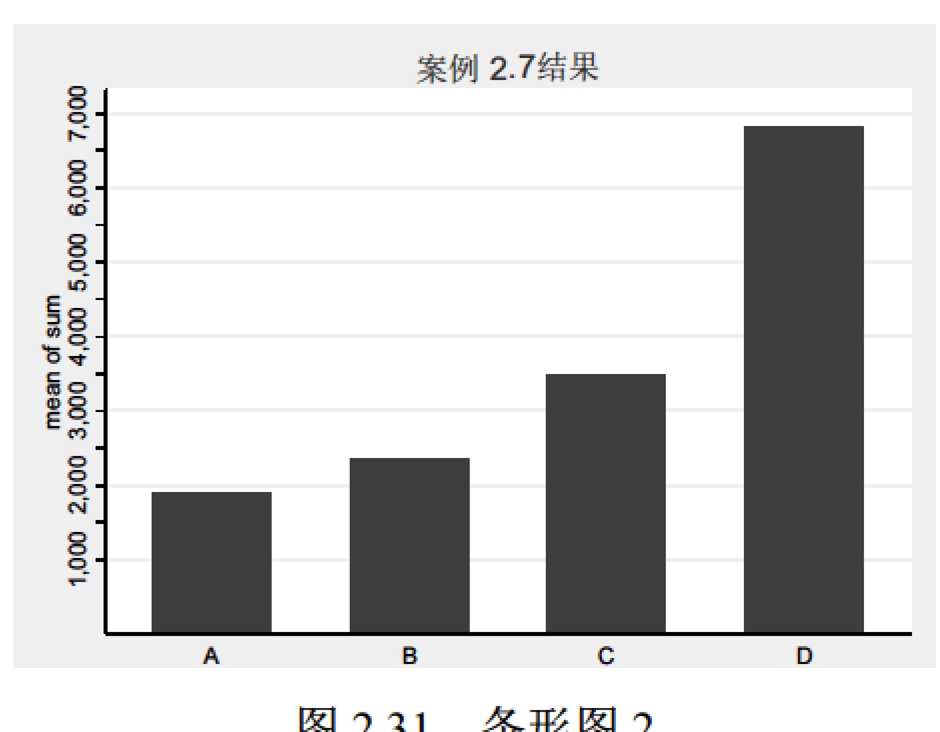


图 2.31　条形图 2

2．延伸 2：利用条形图进行对比分析以得到更多信息

例如，我们要在延伸 1 的基础上对问题进行深入研究，在上面的案例中得到了各团队工作总业绩的具体排名，那么这种总业绩的差异是不是由于团队人数的差异引起的？是否高工作业绩的团队配备了更多的员工？下面我们采用新的命令分析一下。操作命令改为：

```
graph bar  sum  number,over(team) title("案例 2.7 结果") ylabel(1000(1000)7000)
ytick(1000(500)7000)
```

在命令窗口输入命令并按回车键进行确认，结果如图 2.32 所示。

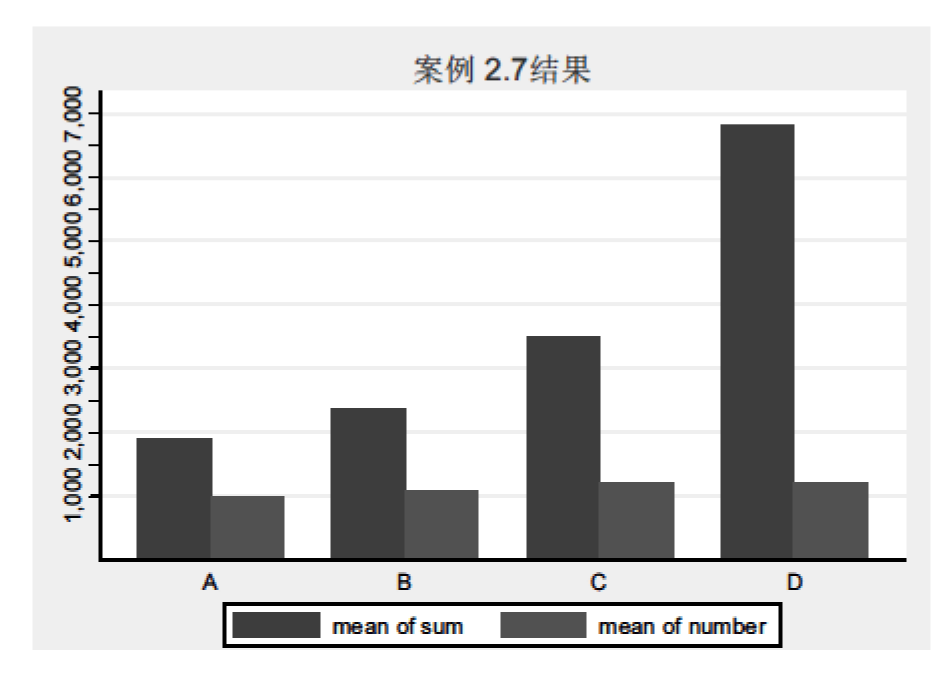


图 2.32　条形图 3

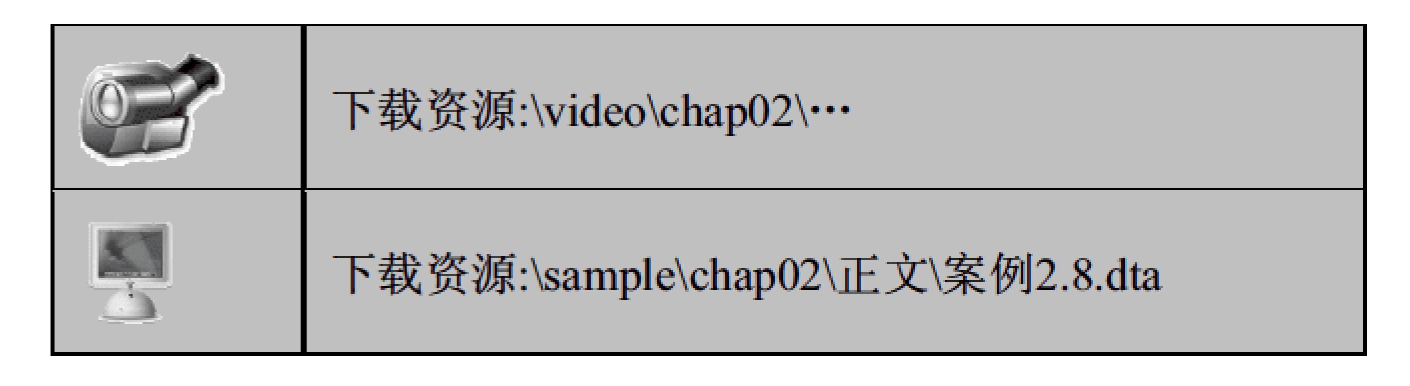
在上面的结果中，我们可以看到该商业银行各团队之间人数的差别是不明显的，也就是说，各团队工作业绩的巨大差别并不是明显地由各团队的员工数量差别引起的。

2.8　实例八——点图

2.8.1　点图的功能与意义

点图（Dot Plot）的功能与作用是和前面提到的条形图类似的，都是用来直观地比较一个或者多个变量的概要统计情况。点图应用广泛，经常出现在政府机关或者咨询机构发布的预测报告中。

2.8.2　相关数据来源



	下载资源:\video\chap02\…
	下载资源:\sample\chap02\正文\案例2.8.dta

【例 2.8】山东省济南市某医院在市内设立有 5 个分院，分别是历下分院、历城分院、天桥分院、槐荫分院、高新分院，以服务各区市民，其内部员工的人数组成如表 2.10 所示。试通过绘制点图按分院分析该医院员工的组成情况。

表 2.10　某医院内部员工人数组成情况

分院名称	男员工人数	女员工人数
历下分院	56	61
历城分院	67	68
天桥分院	66	71
槐荫分院	59	67
高新分院	78	81

2.8.3　Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有 3 个变量，分别是分院名称、男员工人数以及女员工人数。我们把分院名称变量设定为 name，把男员工人数变量设定为 man，把女员工人数变量设定为 woman，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作我们在第 1 章中已有详细讲述。录入完成后数据如图 2.33 所示。

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入命令：

```
graph dot  man wowan,over( name)
```

03 设置完毕后，按键盘上的回车键，等待输出结果。

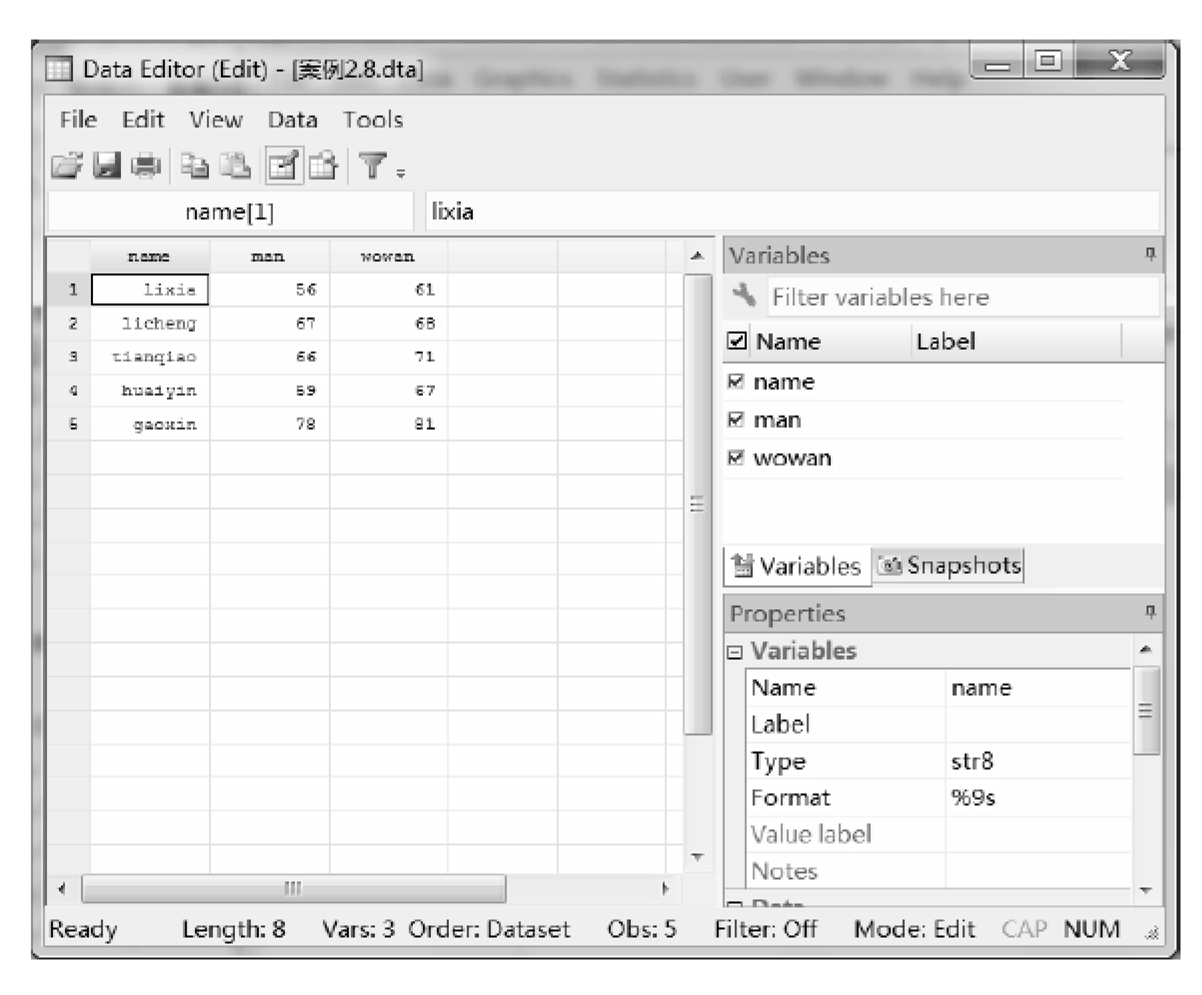


图 2.33　案例 2.8 数据

2.8.4　结果分析

上述操作完成后，Stata 14.0 会弹出如图 2.34 所示的点图。

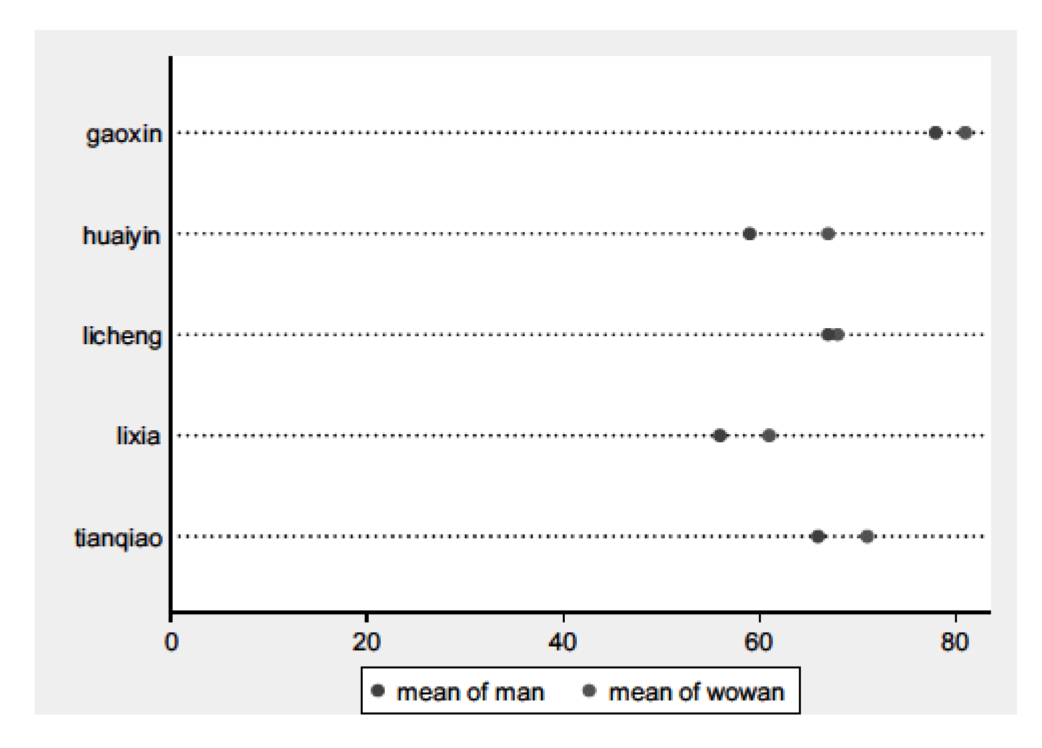


图 2.34　点图 1

通过观察该点图，可以比较轻松地看出很多信息：第一，各个分院的女员工人数都比男员工人数多，因为代表女员工的点都在代表男员工的点的右侧；第二，高新分院不论是男员工还是女员工，人数都是最多的；第三，历下分院不论是男员工还是女员工，人数都是最少的。

2.8.5　案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

1．延伸 1：给图形增加标题

例如，我们要给图形增加标题名称“案例 2.8 结果”，那么操作命令就应该相应地修改为：

```
graph dot  man wowan,over( name) title("案例 2.8 结果")
```

在命令窗口输入命令并按回车键进行确认，结果如图 2.35 所示。

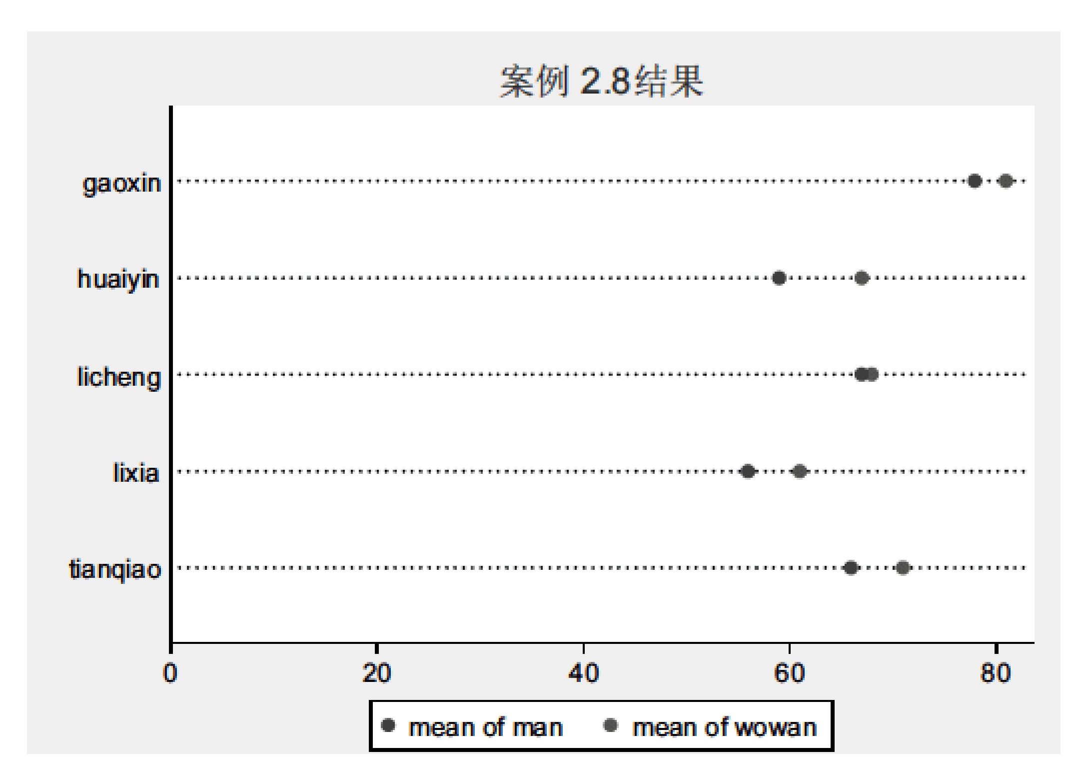


图 2.35　点图 2

2．延伸 2：控制散点标志的形状

此处与散点图略有不同，我们需要用到 marker 命令。例如，我们要在延伸 1 的基础上使图中男性员工散点标志的形状变为实心菱形，使图中女性员工散点标志的形状变为实心三角，那么操作命令就应该相应地修改为：

```
graph dot  man wowan,over( name) title("案例 2.8 结果") marker(1,msymbol(D))
marker(2,msymbol(T))
```

在命令窗口输入命令并按回车键进行确认，结果如图 2.36 所示。

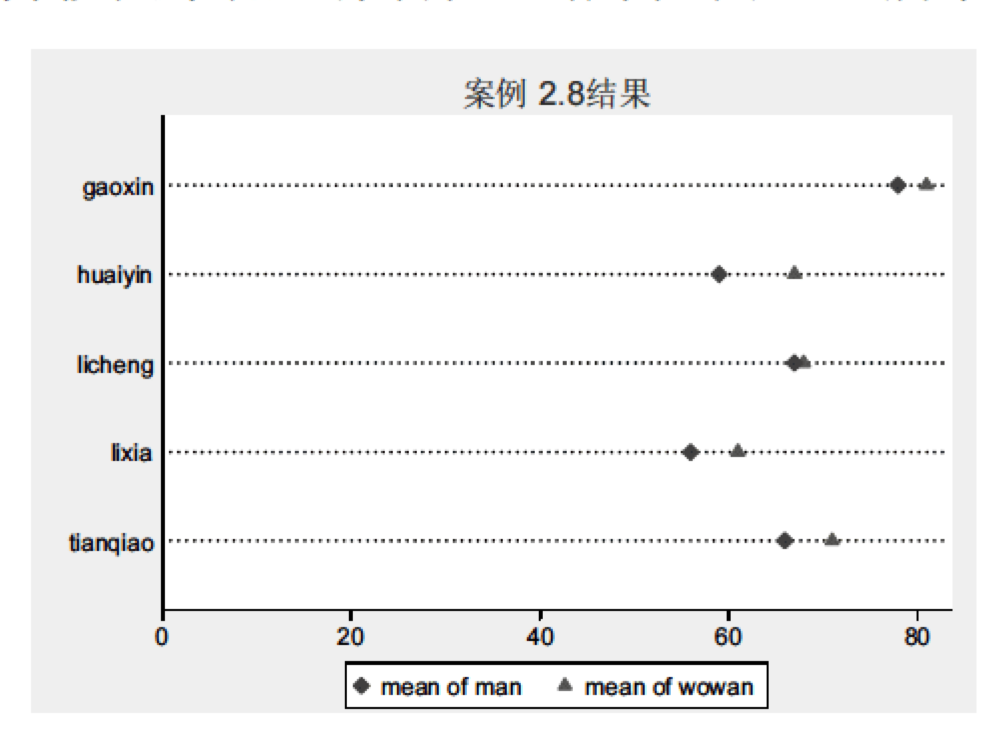


图 2.36　点图 3

2.9　本章习题

（1）为了解我国各地区的电力消费情况，某课题组搜集整理了 2009 年我国 29 个省市的电力消费数据，如表 2.11 所示。试通过绘制直方图来直观地反映我国各地区的电力消费情况。

表 2.11　2009 年我国 29 省市的电力消费情况

地区	电力消费/亿千瓦时
北京	739.146
天津	550.156
河北	2343.85
山西	1267.54
内蒙古	1287.93
…	…
青海	337.237
宁夏	462.958
新疆	547.877

（2）为了解某班级学生的学习情况，教师对该班的学生举行了一次封闭式测验，成绩如表 2.12 所示。试通过绘制散点图来直观地反映这些学生的语文、数学成绩的组合情况。

表 2.12　某班级学生的学习成绩

编号	语文成绩	数学成绩
1	99	67
2	97	77
3	90	77
4	67	59
5	67	64
…	…	…
40	66	89

（续表）

编号	语文成绩	数学成绩
41	63	69
42	60	91

（3）某山村有每年自行进行人口普查的习惯，该山村近些年的人口数据如表 2.13 所示。试通过绘制曲线标绘图来分析研究该山村的人口情况变化趋势以及新生儿对总人口数的影响程度。

表 2.13　某山村人口普查资料

年份	总人数	新生儿数
1997	128	15
1998	138	16
1999	144	16
2000	156	17
2001	166	21
…	…	…
2010	210	39
2011	215	38
2012	219	41

（4）某课题研究组准备对我国上市公司的数量情况进行调查研究，调查得到的数据经整理后如表 2.14 所示。试通过绘制连线标绘图来分析研究我国上市公司数量的变化情况。

表 2.14　我国上市公司数量（1998—2009 年）

年份	上市公司数量
1998	851
1999	949
2000	1088
2001	1160
2002	1224
…	…
2007	1550
2008	1625
2009	1718

（5）T 集团是一家国内大型旅游公司，该公司在组织架构上采取的是事业部制管理方式，把全国各分支机构分为 3 个大区，由各分区督导各省市分公司。T 集团在全国各省市的营业额情况如表 2.15 所示。试绘制箱图来研究分析其分布规律。

表 2.15　T 集团各省市的营业额情况

地区	营业额/万元	所属大区
北京	98	1
天津	64	1
河北	39	1
山西	18	1
内蒙古	69	1
…	…	…
青海	39	3
宁夏	18	3
新疆	69	3

（6）Y 公司是一家饮料代理销售公司，公司销售范围包括可乐、奶茶、牛奶等，公司采取区域事业部制的组织架构，在东部、中部、西部都有自己的分部，较为独立地负责本部各产品的具体运营。该公司各大分部的具体营业收入数据如表 2.16 所示。试通过绘制饼图的方式研究该公司各饮料的销售占比情况。

表 2.16　Y 公司各饮料的销售占比情况

地区	可乐销售收入/万元	奶茶销售收入/万元	牛奶销售收入/万元
东部	1998	10 235	9837
中部	928	7780	6573
西部	361	1098	1076

（7）某集团内设 4 个产品部，分别为 A、B、C、D，其创造利润以及部门人数的具体情况如表 2.17 所示。试通过绘制条形图的方式来对比分析各部门的工作业绩。

表 2.17　某集团各部门的营业净收入情况

产品部	创造利润/万元	部门人数
A	1143	1028
B	1259	1245
C	1359	1241
D	1478	1200

（8）某银行在国内设有 5 家分行，分别是山东分行、陕西分行、山西分行、北京分行、天津分行，以便为广大客户服务，其内部员工人数的组成结构如表 2.18 所示。试通过绘制点图按分行分析该银行员工的组成情况。

表 2.18　某银行内部员工人数组成情况

分行名称	男员工人数	女员工人数
山东分行	138	152
陕西分行	234	259
山西分行	159	186
北京分行	67	99
天津分行	98	108

第 3 章　Stata 描述统计

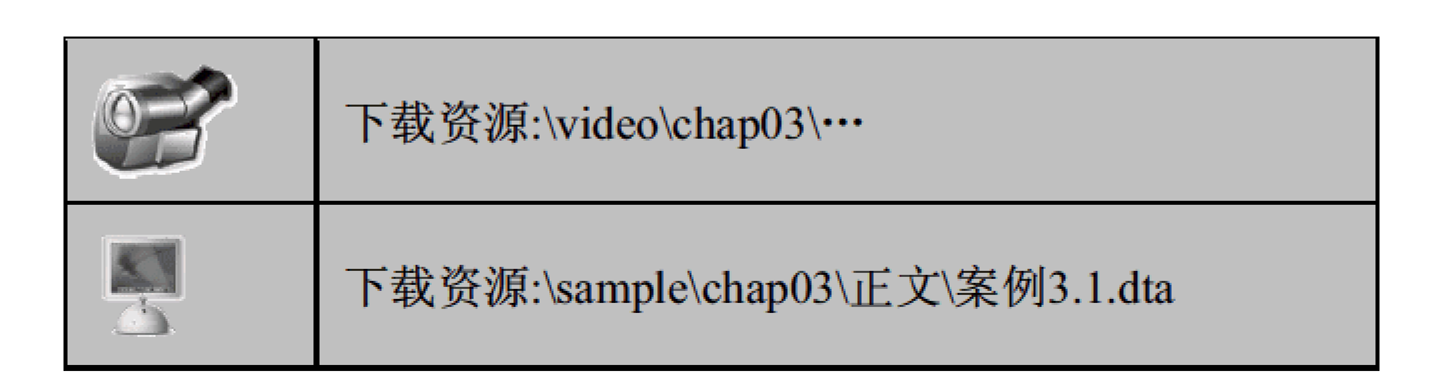
在进行数据分析时，当研究者得到的数据量很小时，可以通过直接观察原始数据来获得所有的信息。但是当得到的数据量很大时，就必须借助各种描述指标来完成对数据的描述工作。用少量的描述指标来概括大量的原始数据，对数据展开描述的统计分析方法被称为描述性统计分析。变量的性质不同，Stata 描述性分析处理的方式也不一样。本章将要介绍的描述统计分析方法包括定距变量的描述性统计、正态性检验和数据转换、单个分类变量的汇总、两个分类变量的列联表分析、多表和多维列联表分析等。下面我们一一介绍这几种方法在实例中的应用。

3.1　实例一——定距变量的描述性统计

3.1.1　定距变量的描述性统计功能与意义

数据分析中的大部分变量都是定距变量，通过进行定距变量的基本描述性统计，我们可以得到数据的概要统计指标，包括平均值、最大值、最小值、标准差、百分位数、中位数、偏度系数和峰度系数等。数据分析者通过获得这些指标，可以从整体上对拟分析的数据进行宏观把握，从而为后续进行更深入的数据分析做好必要的准备。

3.1.2　相关数据来源



	下载资源:\video\chap03\···
	下载资源:\sample\chap03\正文\案例3.1.dta

【例 3.1】为了解我国各地区的电力消费情况，某课题组搜集整理了 2009 年我国 31 个省市的电力消费量的有关数据，如表 3.1 所示。试通过对数据进行基本描述性分析来了解我国各地区电力消费的基本情况。

表 3.1　2009 年我国 31 个省市的电力消费量的有关数据

地区	电力消费量/亿千瓦时
北京	739.146
天津	550.156
河北	2343.85
山西	1267.54
内蒙古	1287.93

（续表）

地区	电力消费量/亿千瓦时
…	…
青海	337.237
宁夏	462.958
新疆	547.877

3.1.3　Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有两个变量，分别是地区和电力消费量。我们把地区变量设定为 region，把电力消费量变量设定为 cunsumption，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作我们在第 1 章中已有详细讲述。录入完成后数据如图 3.1 所示。

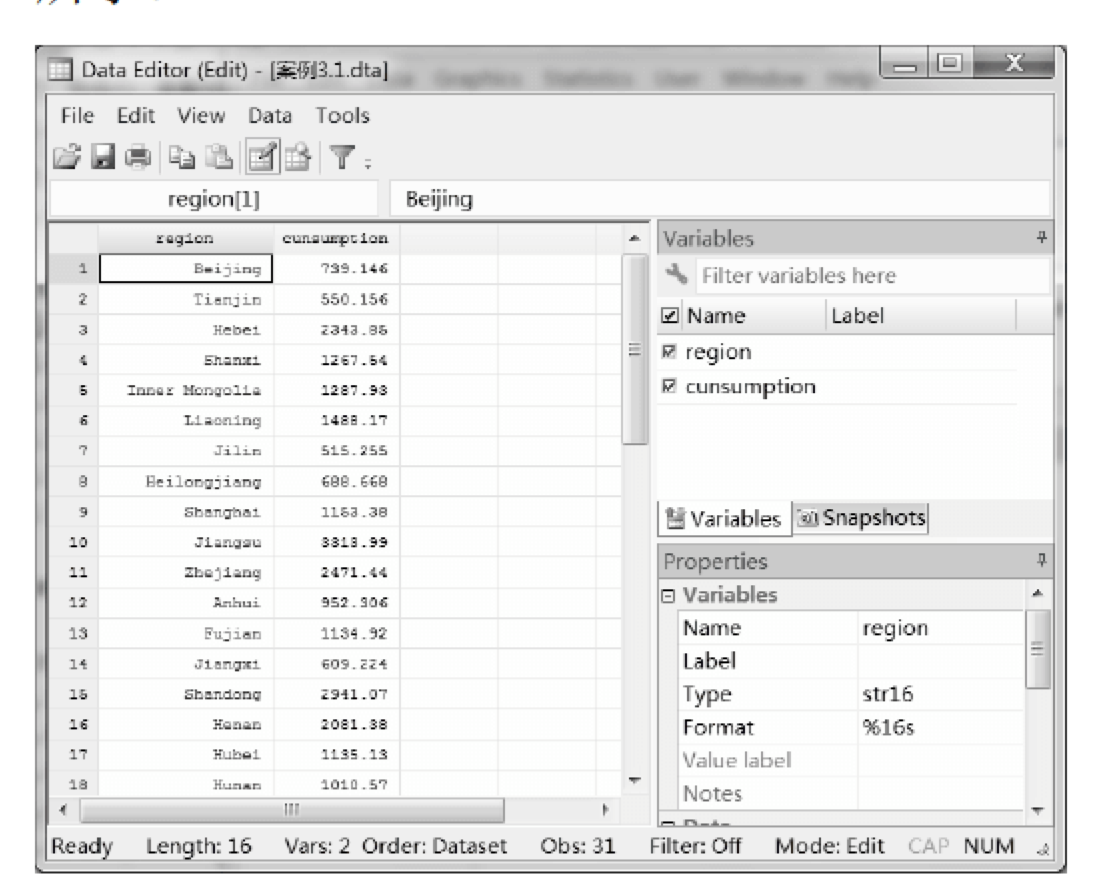


图 3.1　案例 3.1 数据

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入命令：

```
summarize  cunsumption
```

03 设置完毕后，按键盘上的回车键，等待输出结果。

3.1.4　结果分析

在 Stata 14.0 主界面的结果窗口中可以看到如图 3.2 所示的分析结果。

```
. summarize  cunsumption

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
 cunsumption |         31    1180.489    903.5561    17.6987   3609.642
```

图 3.2　分析结果图

通过观察分析结果，我们可以对 2009 年我国各地区的电力消费量情况有一个整体初步的了解。从结果可以看出，有效观测样本共有 31 个，2009 年我国各地区电力消费量的平均值为 1180.489 亿千瓦时，样本的标准差是 903.5561，样本的最小值是 17.6987，样本的最大值是 3609.642。

3.1.5 案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

1．延伸 1：获得更详细的描述性统计结果

操作命令可以相应地修改为：

```
summarize  cunsumption,detail
```

在命令窗口输入命令并按回车键进行确认，结果如图 3.3 所示。

. summarize cunsumption,detail

cunsumption

	Percentiles	Smallest		
1%	17.6987	17.6987		
5%	133.7675	133.7675		
10%	462.9585	337.2368	Obs	31
25%	550.1556	462.9585	Sum of Wgt.	31
50%	891.1902		Mean	1180.489
		Largest	Std. Dev.	903.5561
75%	1324.61	2471.438		
90%	2471.438	2941.067	Variance	816413.7
95%	3313.986	3313.986	Skewness	1.309032
99%	3609.642	3609.642	Kurtosis	3.889152

图 3.3　分析结果图

从上面的分析结果中可以得到更多信息。

（1）百分位数（Percentiles）

可以看出数据的第 1 个四分位数（25%）是 550.1556，数据的第 2 个四分位数（50%）是 891.1902，数据的第 3 个四分位数（75%）是 1324.61。数据的百分位数的含义是低于该数据值的样本在全体样本中的百分比。例如，本例中 25%分位数的含义是全体样本中有 25%的数据值低于 550.1556。

（2）4 个最小值（Smallest）

本例中，最小的 4 个数据值分别是 17.6987、133.7675、337.2368、462.9585。

（3）4 个最大值（Largest）

本例中，最大的 4 个数据值分别是 3609.642、3313.986、2941.067、2471.438。

（4）平均值（Mean）和标准差（Std. Dev）

与前面的分析结果一样，样本数据的平均值为 1180.489，样本数据的标准差是 903.5561。

（5）偏度（Skewness）和峰度（Kurtosis）

偏度的概念是表示不对称的方向和程度。如果偏度值大于 0，那么数据就具有正偏度（右边有尾巴）；如果偏度值小于 0，那么数据就具有负偏度（左边有尾巴）；如果偏度值等于 0，那么数据将呈对称分布。本例中，数据偏度为 1.309032，为正偏度但不大。

峰度的概念用来表示尾重，是与正态分布结合在一起进行考虑的。正态分布是一种对称分布，它的峰度值正好等于 3，如果某数据的峰度值大于 3，那么该分布将会有一个比正态分布更长的尾巴；如果某数据的峰度值小于 3，那么该分布将会有一个比正态分布更短的尾巴。本例中，数据峰度为 3.889152，有一个比正态分布更长的尾巴。

2．延伸 2：根据自己的需要获取相应的概要统计指标

例如，我们想观察各地区电力消费量数据的平均数、总和、极差、方差等数据，那么操作命令可以相应地修改为：

```
tabstat  cunsumption,stats(mean range sum var)
```

在命令窗口输入命令并按回车键进行确认，结果如图 3.4 所示。

```
. tabstat  cunsumption,stats(mean range sum var)

    variable |      mean     range       sum  variance
-------------+----------------------------------------
 cunsumption |  1180.489  3591.944  36595.15  816413.7
```

图 3.4　分析结果图

从上面的分析结果中可以得到更多信息。该样本数据的均值是 1180.489，极差是 3591.944，总和是 36595.15，方差是 816413.7。

统计量与其对应的命令代码如表 3.2 所示。

表 3.2　统计量与其对应的命令代码

统计量	命令代码	统计量	命令代码	统计量	命令代码
均值	mean	非缺失值总数	count	计数	n
总和	sum	最大值	max	最小值	min
极差	range	标准差	sd	方差	var
变异系数	cv	标准误	semean	偏度	skewness
峰度	kurtosis	中位数	median	第1个百分位数	p1
四分位距	iqr	四分位数	q		

3．延伸 3：按另一变量分类列出某变量的概要统计指标

例如，我们要在延伸 2 的基础上按各个省市分别列出数据的概要统计指标，那么操作命令就应该相应地修改为：

```
tabstat  cunsumption,stats(mean range sum var) by(region)
```

在命令窗口输入命令并按回车键进行确认，结果如图 3.5 所示。

```
. tabstat cunsumption,stats(mean range sum var) by(region)

Summary for variables: cunsumption
     by categories of: region
```

region	mean	range	sum	variance
Anhui	952.3056	0	952.3056	.
Beijing	739.1465	0	739.1465	.
Chongqing	533.7976	0	533.7976	.
Fujian	1134.918	0	1134.918	.
Gansu	705.5127	0	705.5127	.
Guangdong	3609.642	0	3609.642	.
Guangxi	856.3511	0	856.3511	.
Guizhou	750.3007	0	750.3007	.
Hainan	133.7675	0	133.7675	.
Hebei	2343.847	0	2343.847	.
Heilongjiang	688.668	0	688.668	.
Henan	2081.375	0	2081.375	.
Hubei	1135.127	0	1135.127	.
Hunan	1010.57	0	1010.57	.
Inner Mongolia	1287.926	0	1287.926	.
Jiangsu	3313.986	0	3313.986	.
Jiangxi	609.2236	0	609.2236	.
Jilin	515.2545	0	515.2545	.
Liaoning	1488.172	0	1488.172	.
Ningxia	462.9585	0	462.9585	.
Qinghai	337.2368	0	337.2368	.
Shaanxi	740.1138	0	740.1138	.
Shandong	2941.067	0	2941.067	.
Shanghai	1153.379	0	1153.379	.
Shanxi	1267.538	0	1267.538	.
Sichuan	1324.61	0	1324.61	.
Tianjin	550.1556	0	550.1556	.
Tibet	17.6987	0	17.6987	.
Xinjiang	547.8766	0	547.8766	.
Yunnan	891.1902	0	891.1902	.
Zhejiang	2471.438	0	2471.438	.
Total	1180.489	3591.944	36595.15	816413.7

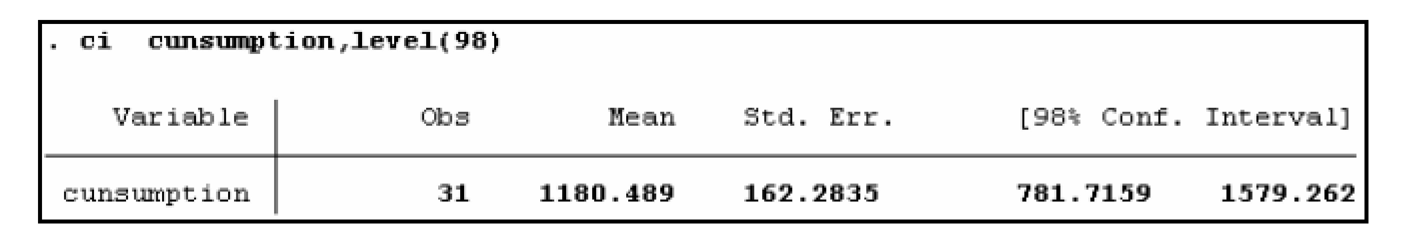
图 3.5　分析结果图

4．延伸 4：创建变量总体均值的置信区间

例如，我们要创建电力消费量均值的 98%的置信区间，那么操作命令就应该相应地修改为：

```
ci  cunsumption,level(98)
```

在命令窗口输入命令并按回车键进行确认，结果如图 3.6 所示。



```
. ci  cunsumption,level(98)
```

Variable	Obs	Mean	Std. Err.	[98% Conf. Interval]	
cunsumption	31	1180.489	162.2835	781.7159	1579.262

图 3.6　分析结果图

基于本例中的观测样本，我们可以推断出总体的 98%水平的置信区间。也就是说，我们有 98%的信心可以认为数据总体的均值会落在[781.7159,1579.262]中，或者说，数据总体的均值落在区间[781.7159,1579.262]的概率是 98%。读者可以根据具体需要通过改变命令中括号里面的数字来调整置信水平的大小。

3.2 实例二——正态性检验和数据转换

3.2.1 正态性检验和数据转换功能与意义

随着科技的不断发展和计算方法的不断改进，学者们探索出了很多统计分析方法和分析程序。但是有相当多的统计程序对数据要求比较严格，它们只有在变量服从或者近似服从正态分布的时候才是有效的，所以在对整理收集的数据进行预处理的时候需要对它们进行正态检验，如果数据不满足正态分布假设，我们就要对数据进行必要的转换。数据转换分为线性转换与非线性转换两种，其中线性转换比较简单，我们在第 1 章中也有所涉及。本节将要讲述的是数据的非线性转换在实例中的应用。

3.2.2 相关数据来源

	下载资源:\video\chap03\···
	下载资源:\sample\chap03\正文\案例3.2.dta

【例 3.2】为了解我国各地区公共交通的运营情况，某课题组搜集整理了我国 2009 年各省市公共交通车辆运营的数据，如表 3.3 所示。试使用 Stata 14.0 对数据进行以下操作：①对该数据进行正态分布检验；②对数据执行平方根变换方法，以获取新的数据并进行正态分布检验；③对数据执行自然对数变换方法，以获取新的数据并进行正态分布检验。

表 3.3 我国 2009 年各省市公共交通车辆运营数据

地区	公共交通车辆运营数/辆
北京	23 730
天津	8 118
河北	13 531
山西	6 655
内蒙古	5 558
…	…
青海	1 994
宁夏	2 133
新疆	8 082

3.2.3 Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有两个变量，分别是地区和公共交通车辆运营数。我们把地区变量设定为 region，把公共交通车辆运营数设定为 sum，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作我们在第 1 章中已有详细讲述。录入完成后数据如图 3.7 所示。

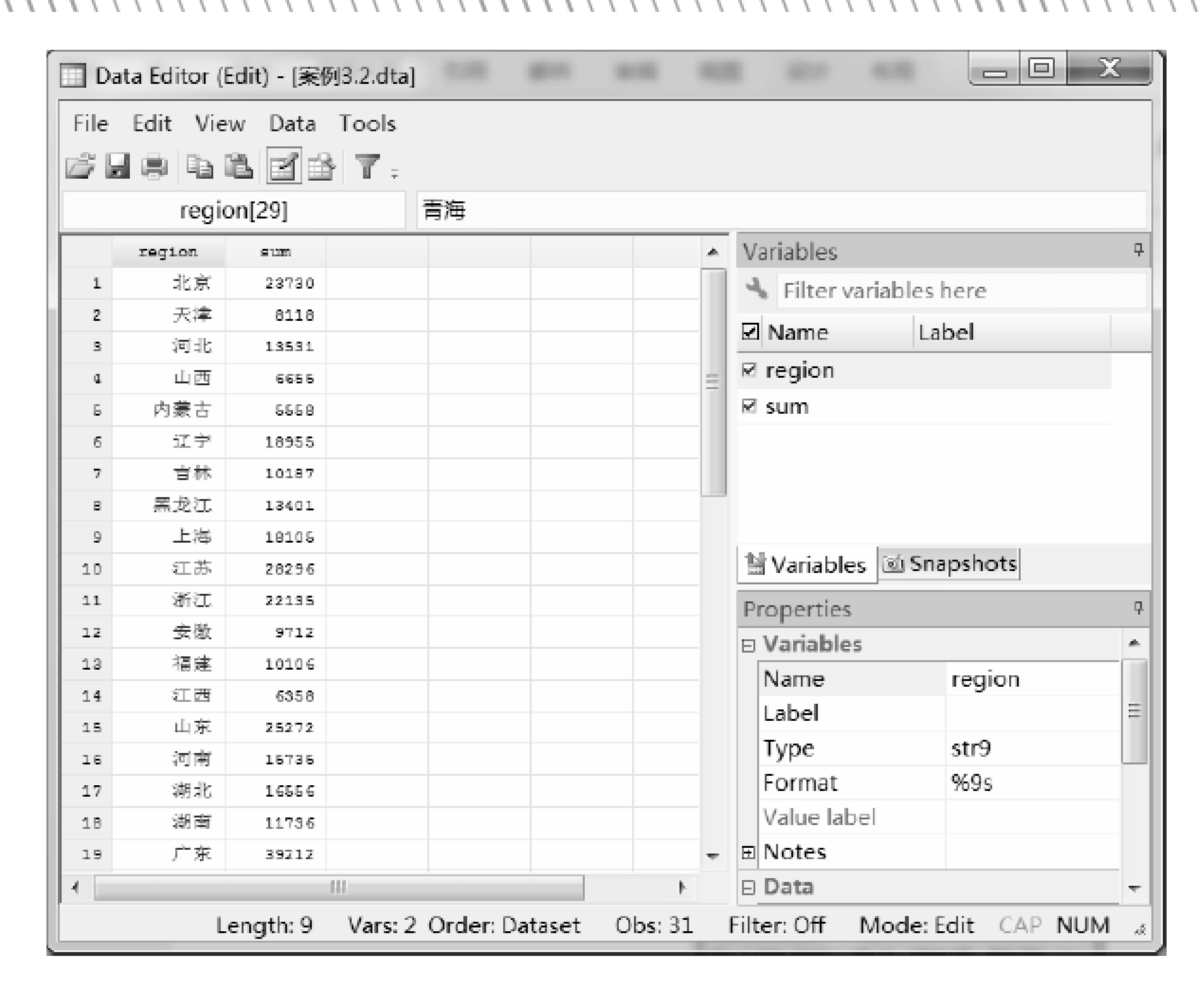


图 3.7　案例 3.2 数据

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

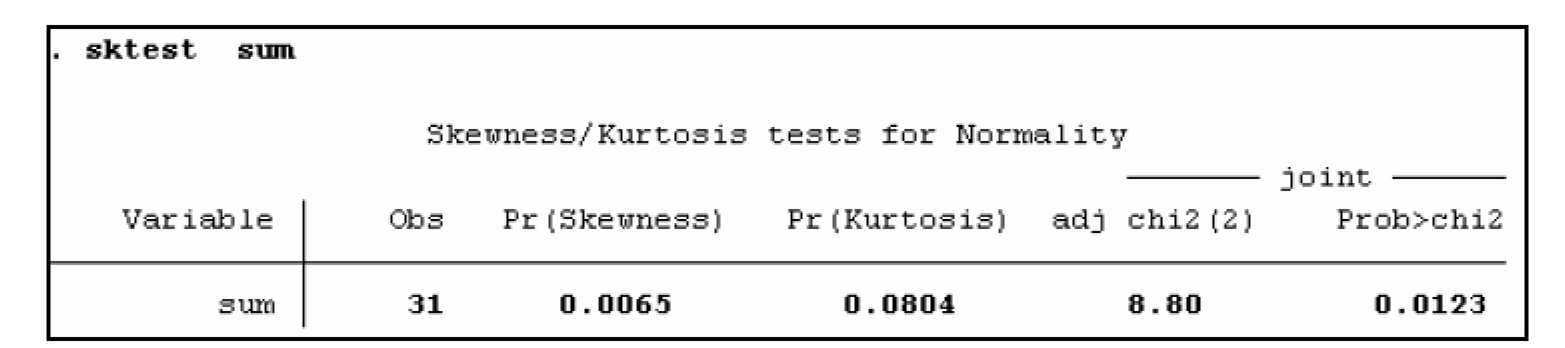
02 在主界面的“Command”文本框中输入操作命令并按键盘上的回车键进行确认。对应的命令分别如下。

- sktest　sum：本命令的含义是对该数据进行正态分布检验。
- generate srsum=sqrt(sum) sktest srsum：本命令的含义是对数据执行平方根变换方法，以获取新的数据并进行正态分布检验。
- generate lsum=ln(sum) sktest lsum：本命令的含义是对数据执行自然对数变换方法，以获取新的数据并进行正态分布检验。

3.2.4　结果分析

在 Stata 14.0 主界面的结果窗口中可以看到如图 3.8~图 3.10 所示的分析结果。

图 3.8 是对该数据进行正态分布检验的结果。



```
. sktest   sum
```

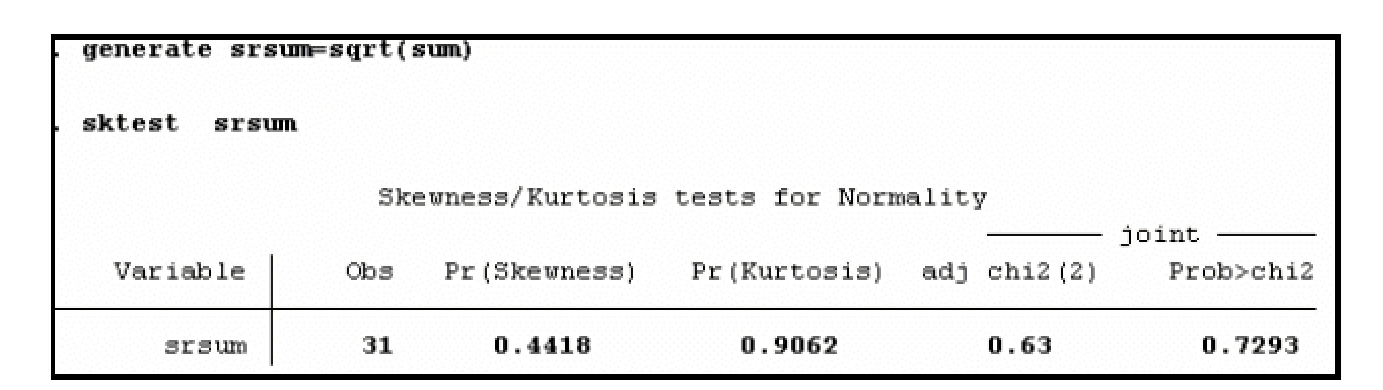
Skewness/Kurtosis tests for Normality

				joint	
Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	adj chi2(2)	Prob>chi2
sum	31	0.0065	0.0804	8.80	0.0123

图 3.8　分析结果图

通过观察分析图，我们可以比较轻松地得出分析结论。本例中，sktest 命令拒绝了数据呈正态分布的原始假设。从偏度上看，Pr(Skewness)为 0.0065，小于 0.05，拒绝正态分布的原假设；从峰度上看，Pr(Kurtosis)为 0.0804，大于 0.05，接受正态分布的原假设；但是把两者结合在一起考虑，从整体上看，Prob>chi2 为 0.0123，小于 0.05，拒绝正态分布的原假设。

图 3.9 是对数据执行平方根变换方法，以获取新的数据并进行正态分布检验的结果。


```
. generate srsum=sqrt(sum)

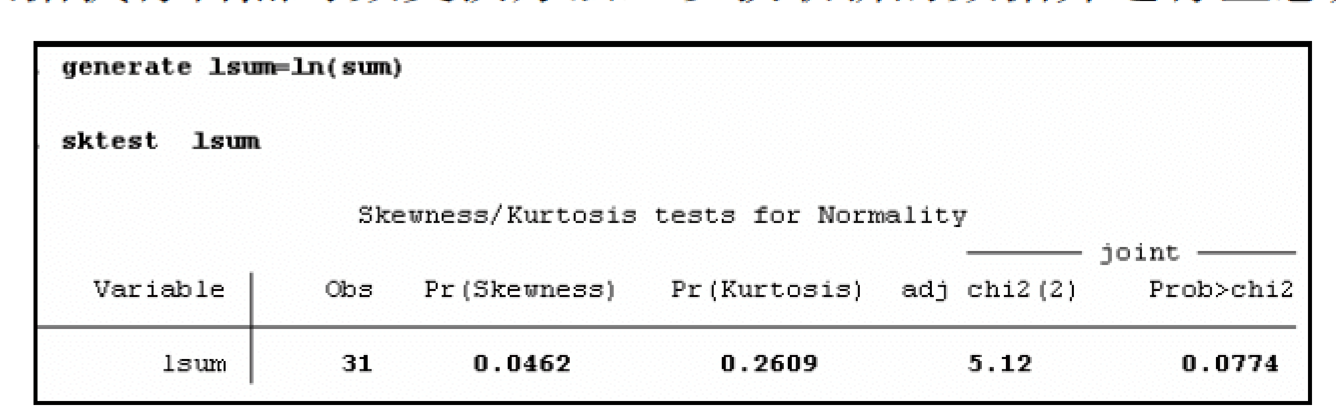
. sktest  srsum

                    Skewness/Kurtosis tests for Normality
                                                          ------- joint ------
    Variable |        Obs  Pr(Skewness)  Pr(Kurtosis) adj chi2(2)   Prob>chi2
-------------+-----------------------------------------------------------------
       srsum |         31     0.4418        0.9062        0.63         0.7293
```

图 3.9　分析结果图

通过观察分析图，我们可以比较轻松地得出分析结论。本例中，sktest 命令接受了数据呈正态分布的原始假设。从偏度上看，Pr(Skewness)为 0.4418，大于 0.05，接受正态分布的原假设；从峰度上看，Pr(Kurtosis)为 0.9062，大于 0.05，接受正态分布的原假设；把两者结合在一起考虑，从整体上看，Prob>chi2 为 0.7293，大于 0.05，接受正态分布的原假设。

图 3.10 是对数据执行自然对数变换方法，以获取新的数据并进行正态分布检验的结果。


```
. generate lsum=ln(sum)

. sktest  lsum

                    Skewness/Kurtosis tests for Normality
                                                          ------- joint ------
    Variable |        Obs  Pr(Skewness)  Pr(Kurtosis) adj chi2(2)   Prob>chi2
-------------+-----------------------------------------------------------------
        lsum |         31     0.0462        0.2609        5.12         0.0774
```

图 3.10　分析结果图

通过观察分析图，我们可以比较轻松地得出分析结论。本例中，sktest 命令接受了数据呈正态分布的原始假设。从偏度上看，Pr(Skewness)为 0.0462，小于 0.05，拒绝正态分布的原假设；从峰度上看，Pr(Kurtosis)为 0.2609，大于 0.05，接受正态分布的原假设；把两者结合在一起考虑，从整体上看，Prob>chi2 为 0.0774，大于 0.05，接受正态分布的原假设。

3.2.5　案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

1．延伸 1：有针对性地对数据进行变换

我们在进行数据分析时，在对初始数据进行正态性检验后，可以利用 3.1 节的相关知识，得到关于数据偏度和峰度的信息，我们完全可以根据数据信息的偏态特征进行有针对性的数据变换。数据变换与其对应的 Stata 命令以及达到的效果如表 3.4 所示。

表 3.4　数据变换与其对应的 Stata 命令以及达到的效果

stata命令	数据转换	效果
generate y=x^3	立方	减少严重负偏态
generate y=x^2	平方	减少轻度负偏态
generate y=sqrt(x)	平方根	减少轻度正偏态
generate y=ln(x)	自然对数	减少轻度正偏态
generate y=log10(old)	以10为底的对数	减少正偏态
generate y=-(sqrt(x))	平方根负对数	减少严重正偏态
generate y=-(x^-1)	负倒数	减少非严重正偏态
generate y=-(x^-2)	平方负倒数	减少非严重正偏态
generate y=-(x^-3)	立方负倒数	减少非严重正偏态

2．延伸 2：关于 ladder 命令的介绍

此处我们介绍一个非常好用的命令：ladder。它把幂阶梯和正态分布检验有效地结合到了一起。它尝试幂阶梯上的每一种幂并逐个反馈结果是否显著地为正态或者非正态。以本例为例，操作命令如下：

```
ladder  sum
```

在命令窗口输入命令并按回车键进行确认，结果如图 3.11 所示。

. ladder sum

Transformation	formula	chi2(2)	P(chi2)
cubic	sum^3	37.26	0.000
square	sum^2	26.32	0.000
identity	sum	8.80	0.012
square root	sqrt(sum)	0.63	0.729
log	log(sum)	5.12	0.077
1/(square root)	1/sqrt(sum)	20.13	0.000
inverse	1/sum	33.29	0.000
1/square	1/(sum^2)	45.24	0.000
1/cubic	1/(sum^3)	47.92	0.000

图 3.11　分析结果图

在该结果中，我们可以非常轻松地看出，在 95%的置信水平上，仅有平方根变换 square root（P(chi2)= 0.729）以及自然对数变换 log（P(chi2)= 0.077）是符合正态分布的，其他幂次的数据变换都不能使数据显著地呈现正态分布。

3．延伸 3：关于 gladder 命令的介绍

例如，我们要在延伸 2 的基础上更直观地看出幂阶梯和正态分布检验有效结合的结果，那么操作命令就应该相应地修改为：

```
gladder  sum
```

在命令窗口输入命令并按回车键进行确认，结果如图 3.12 所示。

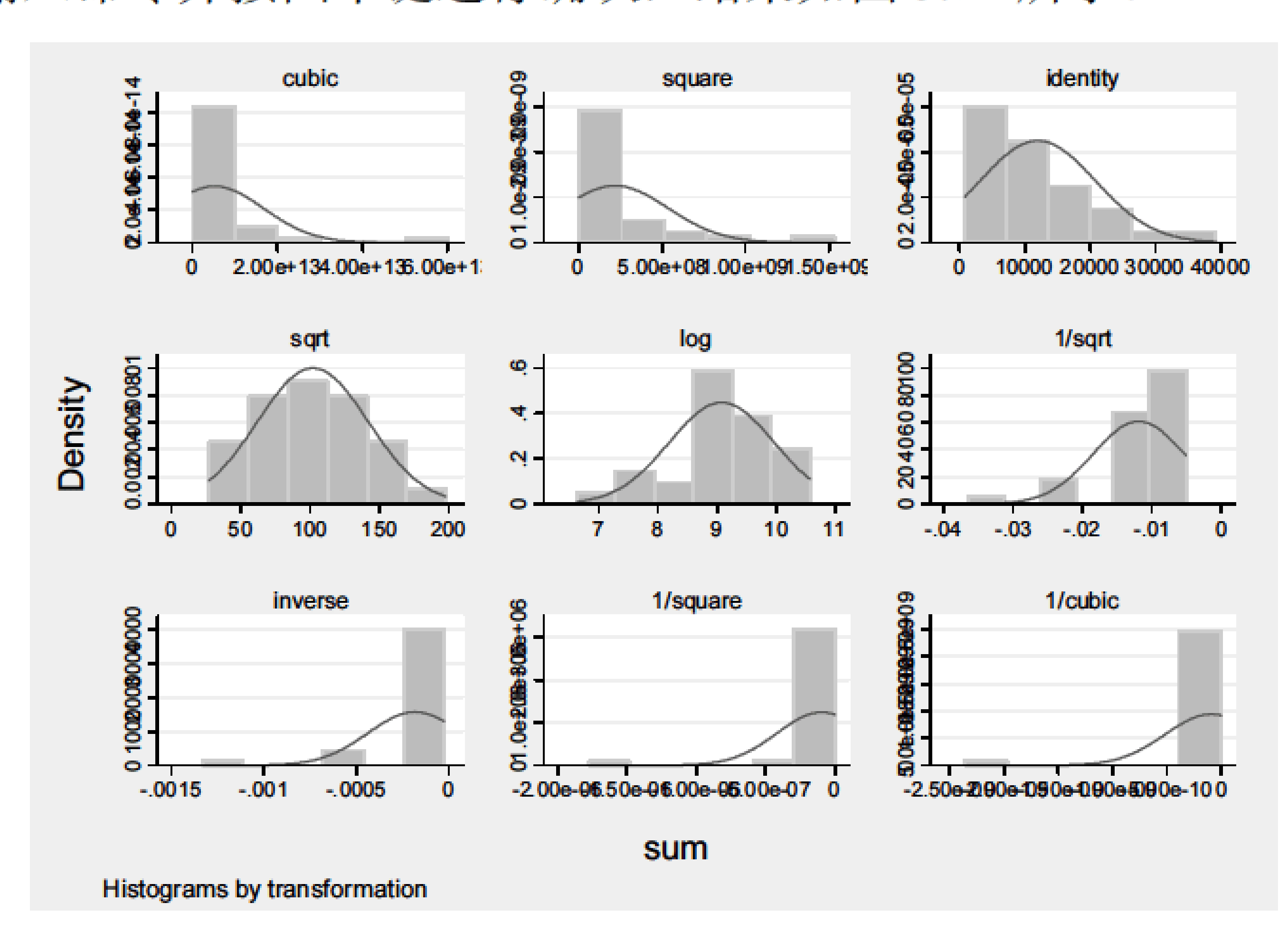


图 3.12　分析结果图

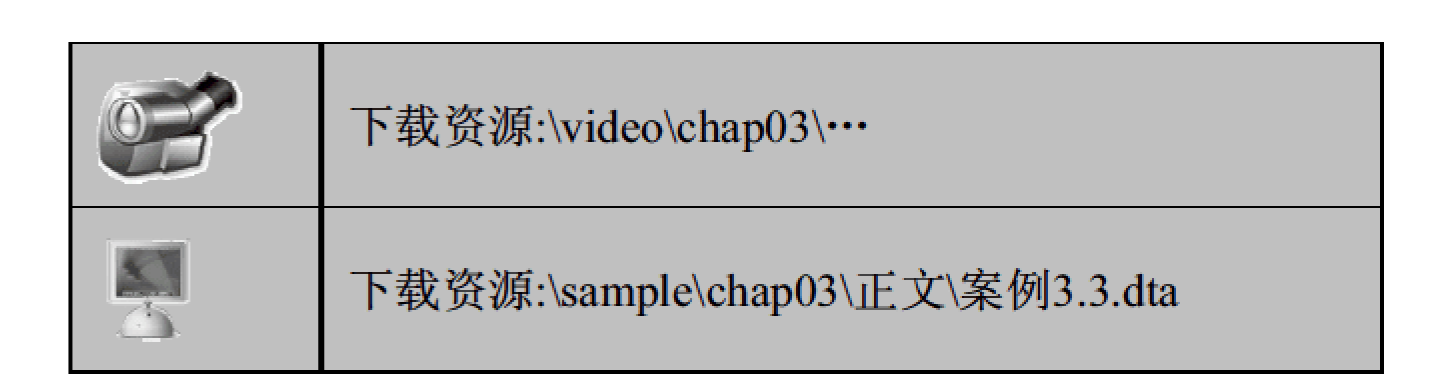
从结果中可以轻松地看出每种转换的直方图与正态分布曲线，与延伸 2 得出的结论是一致的。

3.3　实例三——单个分类变量的汇总

3.3.1　单个分类变量的汇总功能与意义

与前面提到的定距变量不同，分类变量的数值只代表观测值所属的类别，不代表其他任何含义。因此，对分类变量的描述统计方法是观察其不同类别的频数或者百分数。本节我们将介绍单个分类变量的汇总在实例中的应用。

3.3.2　相关数据来源



下载资源:\video\chap03\…

下载资源:\sample\chap03\正文\案例3.3.dta

【例 3.3】某国有银行沈阳分行人力资源部对分行本部在岗职工的结婚情况进行了调查。调查结果分为了两类，一类代表结婚，另一类代表未婚或者离异。统计数据如表 3.5 所示。试对结婚情况这一变量进行单个变量汇总。

表 3.5　某银行沈阳分行本部在岗职工的结婚情况

编号	性别	结婚情况
1	女	是
2	男	是
3	男	是
4	男	否
5	男	是
…	…	…
112	女	是
113	男	是
114	女	否

3.3.3　Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有两个变量，分别为性别和结婚情况。我们把性别变量设定为 gender，把结婚情况变量设定为 marry，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作我们在第 1 章中已有详细讲述。录入完成后数据如图 3.13 所示。

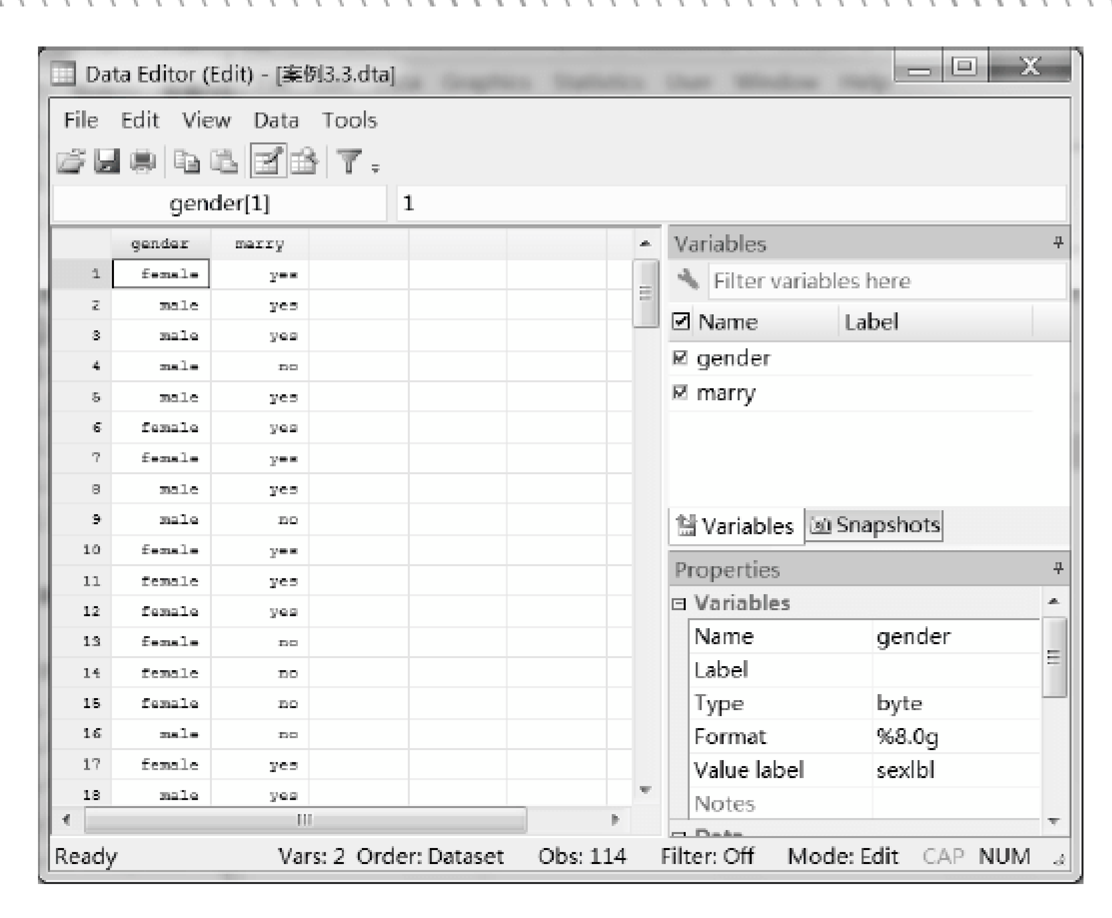


图 3.13　案例 3.3 数据

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入命令：

```
tabulate marry
```

03 设置完毕后，按键盘上的回车键，等待输出结果。

3.3.4　结果分析

在 Stata 14.0 主界面的结果窗口中可以看到如图 3.14 所示的分析结果。

```
. tabulate  marry

      marry |      Freq.     Percent        Cum.
------------+-----------------------------------
         no |         45       39.47       39.47
        yes |         69       60.53      100.00
------------+-----------------------------------
      Total |        114      100.00
```

图 3.14　曲线标绘图 1

从分析结果中我们可以看出本次调查所获得的信息。可以发现该银行的分行本部共有 114 人参与了有效调查，其中处于结婚状态的有 69 位员工，占比 60.53%，处于非结婚状态的有 45 位员工，占比 39.47%。此外，结果分析表中 Cum.一栏表示的是累计百分比。

3.3.5　案例延伸

以本节所介绍的案例为基础，试对结婚情况这一变量进行单个变量汇总并附有星点图。操作命令应该为：

```
tabulate marry,plot
```

在命令窗口输入命令并按回车键进行确认，结果如图 3.15 所示。

```
. tabulate marry,plot

      marry |      Freq.
------------+------------+-----------------------------------------------------
         no |         45 |*************************************
        yes |         69 |*********************************************************
------------+------------+-----------------------------------------------------
      Total |        114
```

图 3.15　分析结果图

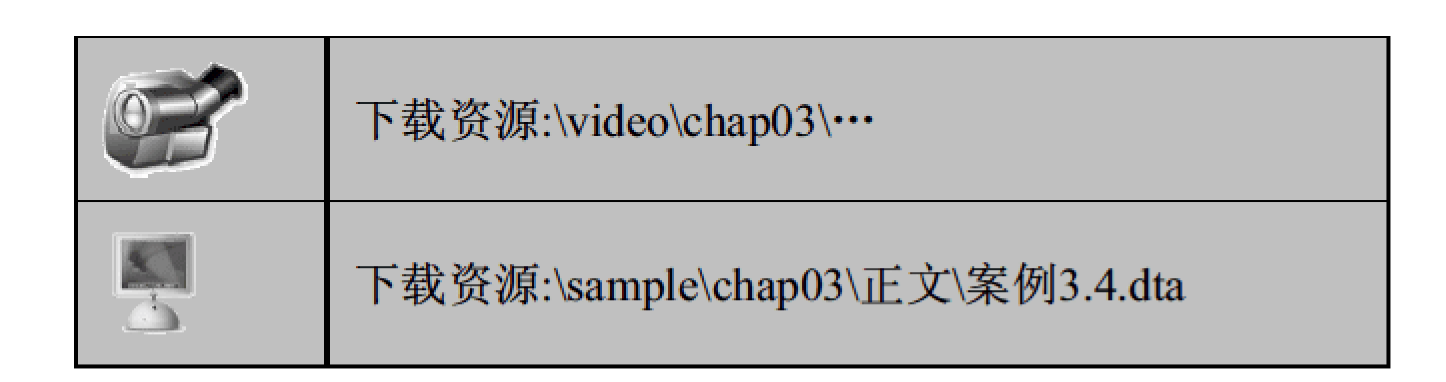
从分析结果中我们可以看出对结婚情况这一变量进行单个变量汇总的结果以及星点图情况。

3.4　实例四——两个分类变量的列联表分析

3.4.1　两个分类变量的列联表分析功能与意义

在上节中，我们讲述了单个分类变量进行概要统计的实例，在本节中，我们将以实例的方式讲解一下两个分类变量是如何进行概要统计的，即二维列联表。

3.4.2　相关数据来源



下载资源:\video\chap03\…

下载资源:\sample\chap03\正文\案例3.4.dta

【例 3.4】为研究 A 市居民的身体情况，某课题组对 A 市居民的吸烟喝酒情况进行了调查研究，调查得到的数据经整理后如表 3.6 所示。试对该数据资料进行二维列联表分析。

表 3.6　A 市居民的吸烟喝酒情况

编号	性别	是否吸烟	是否喝酒
1	女	否	否
3	女	否	是
3	女	否	否
4	男	是	是
5	男	否	是
…	…	…	…
122	女	是	是
123	男	否	否
124	男	是	是

3.4.3　Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。容易发现本例中有 3 个变量，分别是性别、是否吸烟以及是否喝酒。我们把性别变量设定为 gender，把是否吸烟变量设定为

smoke，把是否喝酒变量设定为 drink，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作我们在第 1 章中已有详细讲述。录入完成后数据如图 3.16 所示。

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入命令：

```
tabulate  smoke drink
```

03 设置完毕后，按键盘上的回车键，等待输出结果。

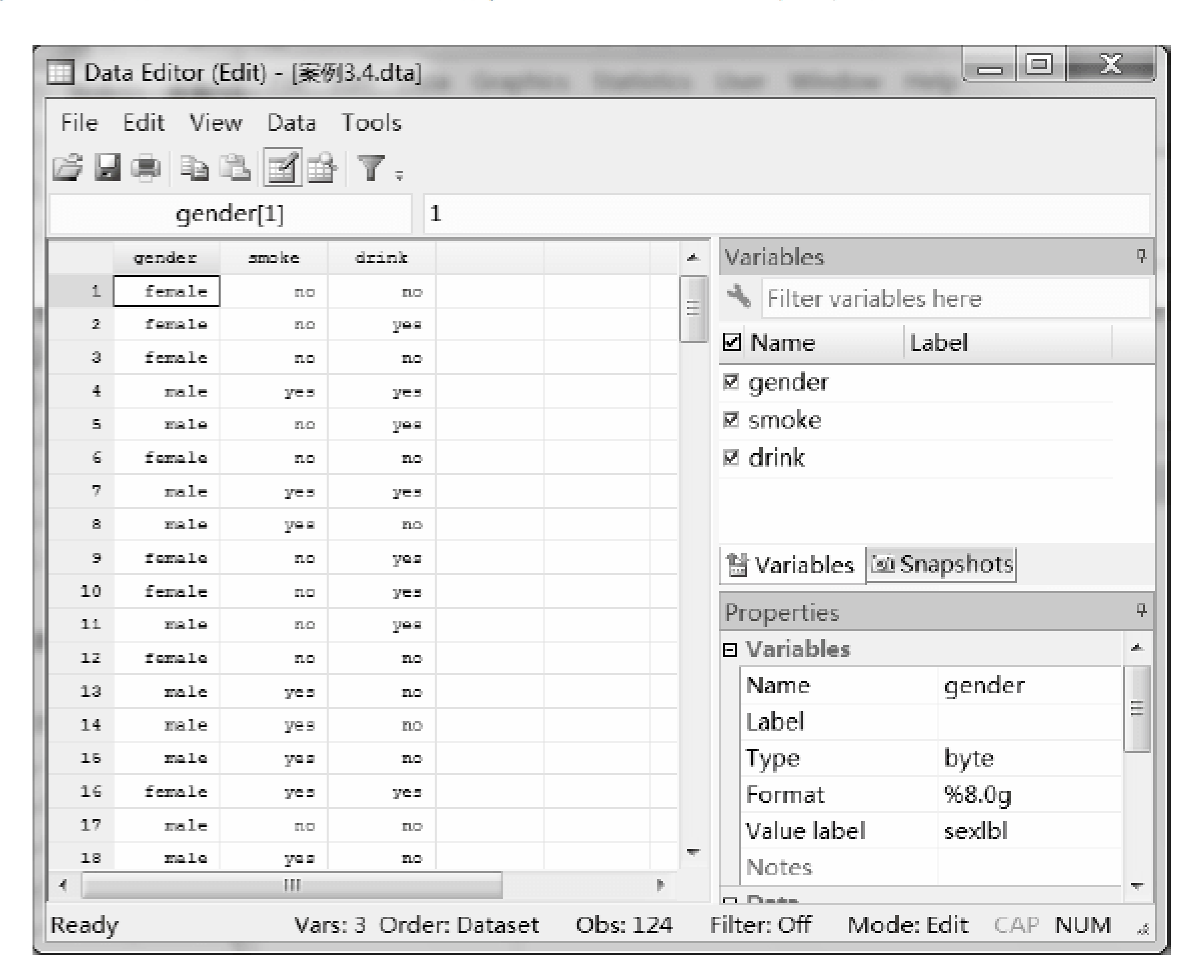


图 3.16　案例 3.4 数据

3.4.4　结果分析

在 Stata 14.0 主界面的结果窗口我们可以看到如图 3.17 所示的分析结果。

```
. tabulate  smoke drink

           |        drink
     smoke |        no        yes |     Total
-----------+----------------------+----------
        no |        44         12 |        56
       yes |        39         29 |        68
-----------+----------------------+----------
     Total |        83         41 |       124
```

图 3.17　分析结果图

从分析结果中可以看出本次调查所获得的信息：发现共有 124 位 A 市居民参与了有效调查，其中有 68 人吸烟，有 56 人不吸烟，有 41 人喝酒，有 83 人不喝酒，具体来说，既吸烟又喝酒的居民人数为 29 人，不吸烟也不喝酒的居民人数为 44 人，只吸烟不喝酒的居民人数为 39 人，只喝酒不吸烟的居民人数为 12 人。

3.4.5 案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

延伸：显示每个单元格的列百分比与行百分比

在本节的例子中，操作命令应该相应地修改为：

```
tabulate  smoke drink,column row
```

在命令窗口输入命令并按回车键进行确认，结果如图 3.18 所示。

```
. tabulate  smoke drink,column row

+-------------------+
| Key               |
|-------------------|
|     frequency     |
|  row percentage   |
| column percentage |
+-------------------+

           |        drink
     smoke |        no        yes |     Total
-----------+----------------------+----------
        no |        44         12 |        56 
           |     78.57      21.43 |    100.00 
           |     53.01      29.27 |     45.16 
-----------+----------------------+----------
       yes |        39         29 |        68 
           |     57.35      42.65 |    100.00 
           |     46.99      70.73 |     54.84 
-----------+----------------------+----------
     Total |        83         41 |       124 
           |     66.94      33.06 |    100.00 
           |    100.00     100.00 |    100.00 
```

图 3.18 分析结果图

分析结果表中的单元格包括 3 部分信息，其中第 1 行表示的是频数，第 2 行表示的是行百分比，第 3 行表示的是列百分比。例如，最左上角的单元格的意义是：不吸烟也不喝酒的样本个数有 44 个，这部分样本在所有不吸烟的样本中占比为 78.57%、在所有不喝酒的样本中占比为 53.01%。

3.5 实例五——多表和多维列联表分析

3.5.1 多表和多维列联表分析功能与意义

对于一些大型数据集，我们经常需要许多不同变量的频数分布。那么如何快速简单地实现这一目的呢？这就需要用到 Stata 的多表和多维列联表分析功能。下面我们就以实例的方式来介绍这一强大功能。

3.5.2 相关数据来源

	下载资源:\video\chap03\…
	下载资源:\sample\chap03\正文\案例3.5.dta

【例 3.5】某高校经济学院针对其研究生学生的持有证书情况进行了调查。证书分为 3 类，包括会计师证书、审计师证书、经济师证书。数据经整理汇总后如表 3.7 所示。试使用 Stata 14.0 对数据进行以下操作：①对数据中的所有分类变量进行单个变量汇总统计；②对数据中的所有分类变量进行二维列联表分析；③以是否持有会计师证书为主分类变量，制作 3 个分类变量的三维列联表。

表 3.7 某高校经济学院的研究生学生持有证书情况

编号	性别	是否持有会计师证书	是否持有审计师证书	是否持有经济师证书
1	男	有	有	无
2	男	有	无	无
3	女	有	有	有
4	女	无	有	有
5	男	无	无	有
…	…	…	…	…
97	女	无	无	无
98	女	有	有	有
99	女	有	有	无

3.5.3 Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有 4 个变量，分别是性别、是否持有会计师证书、是否持有审计师证书以及是否持有经济师证书。我们把性别变量设定为 gender，把是否持有会计师证书设定为 account，把是否持有审计师证书设定为 audit，把是否持有经济师证书设定为 economy，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作我们在第 1 章中已有详细讲述。录入完成后数据如图 3.19 所示。

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入操作命令并按键盘上的回车键进行确认。对应的命令分别如下。

- tab1 account audit economy：本命令的含义是对数据中的所有分类变量进行单个变量汇总统计。

- tab2 account audit economy：本命令的含义是对数据中的所有分类变量进行二维列联表分析。
- by account,sort:tabulate audit economy：本命令的含义是以是否持有会计师证书为主分类变量，制作 3 个分类变量的三维列联表。

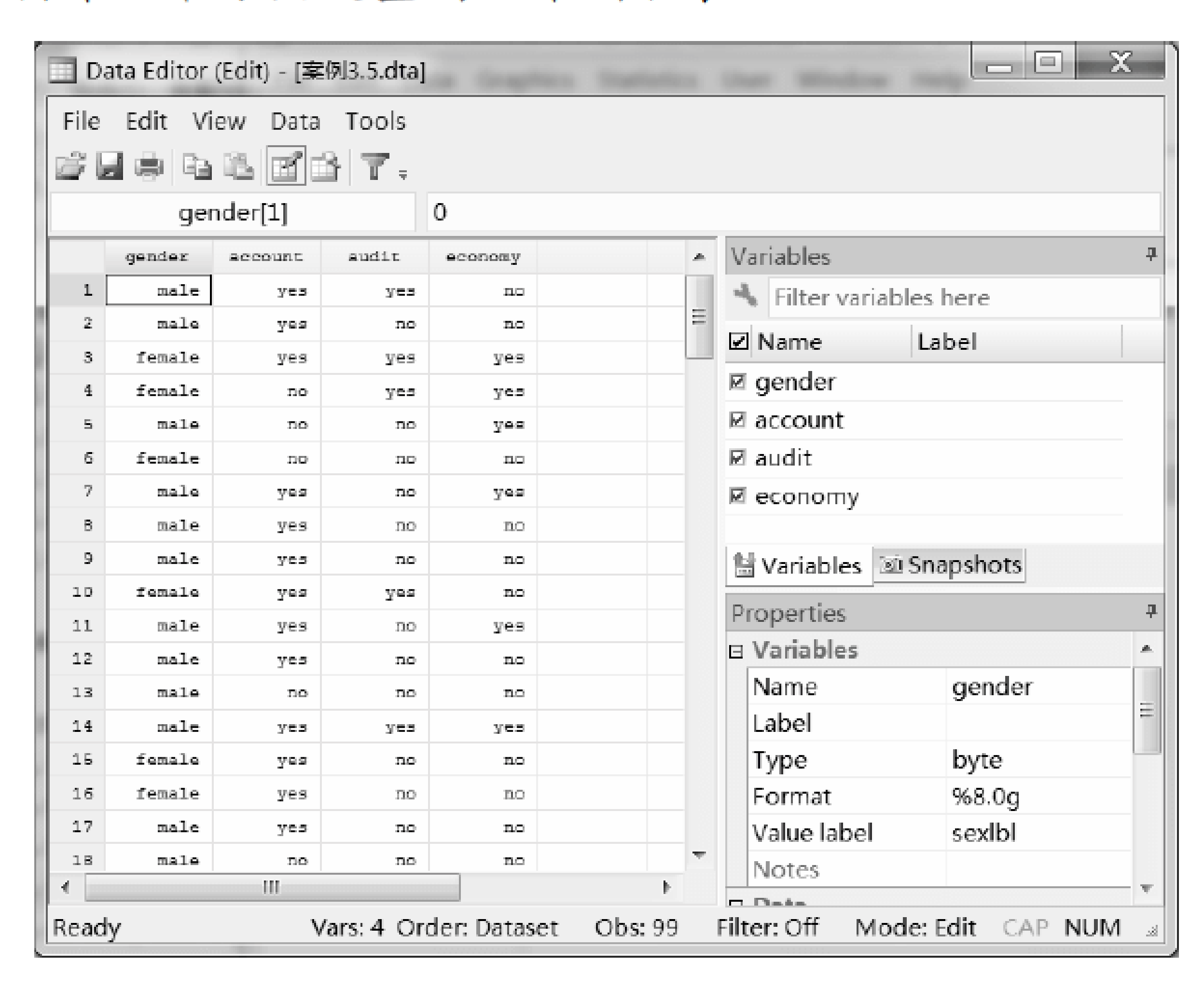
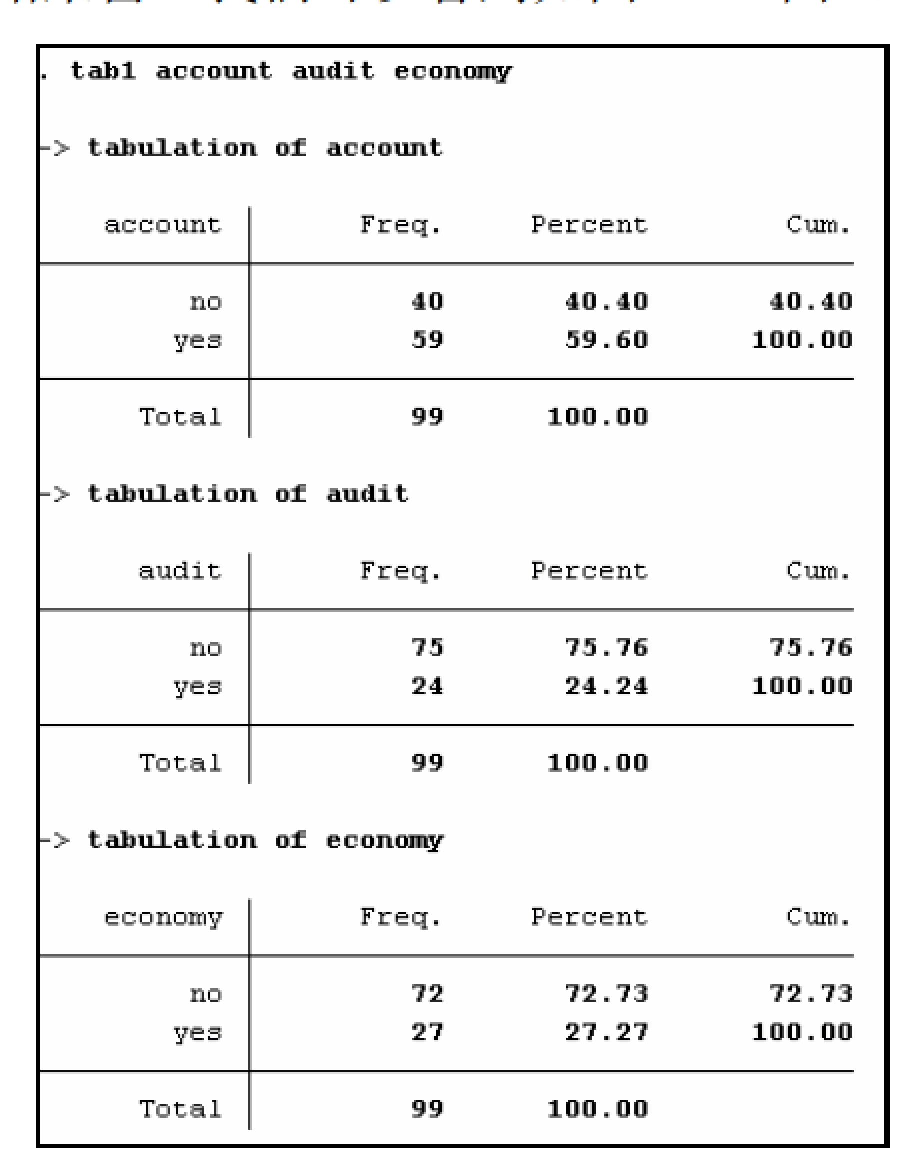


图 3.19 案例 3.5 数据

3.5.4 结果分析

在 Stata 14.0 主界面的结果窗口我们可以看到如图 3.20~图 3.22 所示的分析结果。



```
. tab1 account audit economy

-> tabulation of account

    account |      Freq.     Percent        Cum.
------------+-----------------------------------
         no |         40       40.40       40.40
        yes |         59       59.60      100.00
------------+-----------------------------------
      Total |         99      100.00

-> tabulation of audit

      audit |      Freq.     Percent        Cum.
------------+-----------------------------------
         no |         75       75.76       75.76
        yes |         24       24.24      100.00
------------+-----------------------------------
      Total |         99      100.00

-> tabulation of economy

    economy |      Freq.     Percent        Cum.
------------+-----------------------------------
         no |         72       72.73       72.73
        yes |         27       27.27      100.00
------------+-----------------------------------
      Total |         99      100.00
```

图 3.20 分析结果图

图 3.20 是对数据中的所有分类变量进行单个变量汇总统计的结果。

从分析结果中我们可以看出本次调查所获得的信息：发现该学校经济学院的研究生学生

中共有 99 人参与了有效调查，其中拥有会计师证书的有 59 位学生，在 99 名学生中占比 59.6%；拥有审计师证书的有 24 位学生，在 99 名学生中占比 24.24%；拥有经济师证书的有 27 位学生，在 99 名学生中占比 27.27%。此外，结果分析表中 Cum.一栏表示的是累计百分比。

图 3.21 是对数据中的所有分类变量进行二维列联表分析的结果。

```
. tab2 account audit economy

-> tabulation of account by audit

           |         audit
   account |        no        yes |     Total
-----------+----------------------+----------
        no |        32          8 |        40
       yes |        43         16 |        59
-----------+----------------------+----------
     Total |        75         24 |        99

-> tabulation of account by economy

           |        economy
   account |        no        yes |     Total
-----------+----------------------+----------
        no |        30         10 |        40
       yes |        42         17 |        59
-----------+----------------------+----------
     Total |        72         27 |        99

-> tabulation of audit by economy

           |        economy
     audit |        no        yes |     Total
-----------+----------------------+----------
        no |        60         15 |        75
       yes |        12         12 |        24
-----------+----------------------+----------
     Total |        72         27 |        99
```

图 3.21　分析结果图

从分析结果中我们可以看出本次调查所获得的信息：分析结果中包括 3 张二维列联表，第 1 张是变量“audit”与变量“account”的二维列联分析，第 2 张是变量“economy”与变量“account”的二维列联分析，第 3 张是变量“audit”与变量“economy”的二维列联分析。关于二维列联表的解读，我们在上节的实例中已经讲述过，不再赘述。

图 3.22 是以是否持有会计师证书为主分类变量，制作 3 个分类变量的三维列联表的结果。

```
. by account,sort:tabulate  audit economy

-> account = no

           |        economy
     audit |        no        yes |     Total
-----------+----------------------+----------
        no |        26          6 |        32
       yes |         4          4 |         8
-----------+----------------------+----------
     Total |        30         10 |        40

-> account = yes

           |        economy
     audit |        no        yes |     Total
-----------+----------------------+----------
        no |        34          9 |        43
       yes |         8          8 |        16
-----------+----------------------+----------
     Total |        42         17 |        59
```

图 3.22　分析结果图

该分析结果是一张三维列联表，包括两部分：上半部分描述的是当“account”变量取值

为“no”的时候，变量“audit”与变量“economy”的二维列联分析；下半部分描述的是当“account”变量取值为“yes”的时候，变量“audit”与变量“economy”的二维列联分析。

3.5.5　案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

在这里我们介绍一个用于多维列联分析的 Stata 命令——table。这是一个多功能的命令，可以实现多种数据的频数、标准差数据特征的列联分析。例如，我们要进行简单的频数列联分析，那么操作命令就应该相应地修改为：

```
table account audit economy,contents(freq)
```

在命令窗口输入命令并按回车键进行确认，结果如图 3.23 所示。

```
. table account audit economy,contents(freq)

----------------------------------------
          |      economy and audit
          | ---- no ----    ---- yes ---
  account |   no    yes       no    yes
----------+-----------------------------
       no |   26      4        6      4
      yes |   34      8        9      8
----------------------------------------
```

图 3.23　分析结果图

本结果分析图的解读方式与前面类似，这里不再赘述。

上述命令中 contents 括号里的内容表示的是频数，该括号内支持的内容与命令符号的对应关系如表 3.8 所示。

表 3.8　contents 括号里支持的内容与命令符号的对应关系

命令符号	括号内支持的内容	命令符号	括号内支持的内容
freq	频数	min x	x的最小值
sd x	x的标准差	median x	x的中位数
count x	x非缺失观测值的计数	mean x	x的平均数
n x	x非缺失观测值的计数	rawsum x	忽略任意规定权数的总和
max x	x的最大值	iqr x	x的四分位距
sum x	x的总和	p1 x	x的第1个百分位数

3.6　本章习题

（1）为了解我国各地区的运营线路网的长度情况，某课题组搜集整理了 2009 年我国 31 个省市的运营线路网长度的有关数据，如表 3.9 所示。试通过对数据进行基本描述性分析来了解我国 31 个省市的运营线路网长度的基本情况。

表 3.9　2009 年我国 31 个省市的运营线路网长度的有关数据

地区	运营线路网长度/千米
北京	228
天津	759
河北	8 410
山西	8 710
内蒙古	2 810
…	…
青海	1 057
宁夏	2 708
新疆	4 241

（2）为了解我国各地区公共交通运营情况，某课题组搜集整理了我国 2009 年各省市出租车辆运营的数据，如表 3.10 所示。试使用 Stata 14.0 对数据进行以下操作：①对该数据进行正态分布检验；②对数据执行平方根变换方法，以获取新的数据并进行正态分布检验；③对数据执行自然对数变换方法，以获取新的数据并进行正态分布检验。

表 3.10　我国 2009 年各省市出租车辆运营数据

地区	年末出租车辆运营数/辆
北京	66 646
天津	31 940
河北	46 597
山西	28 729
内蒙古	43 084
…	…
青海	7 041
宁夏	12 582
新疆	24 650

（3）某会计师事务所针对其员工 CPA 证书的持证情况进行了调查。调查结果分为两类：一类代表通过 CPA 考试；另一类代表未通过 CPA 考试。统计数据如表 3.11 所示。试对是否通过 CPA 考试这一变量进行单个变量汇总。

表 3.11　某会计师事务所在岗员工 CPA 证书的持证情况

编号	性别	通过CPA考试情况
1	男	否
2	女	是
3	女	是
4	男	是
5	女	否
…	…	…
127	男	否
128	女	是
129	女	是

（4）某企业面临经营困境，准备进行深刻而彻底的变革。在变革前其对企业员工针对降

薪、降级情况进行了调查研究，调查得到的数据经整理后如表 3.12 所示。试对该数据资料进行二维列联表分析。

表 3.12　某企业员工针对改革措施的看法

编号	性别	是否支持降薪决定	是否支持降级决定
1	女	是	是
3	女	是	是
3	女	是	否
4	男	是	否
5	男	是	否
…	…	…	…
101	女	是	否
102	男	是	否
103	女	否	否

（5）某艺术学校针对其学生的特长情况进行了调查。特长分为 3 类，包括音乐、体育、美术。数据经整理汇总后如表 3.13 所示。试使用 Stata 14.0 对数据进行以下操作：①对数据中的所有分类变量进行单个变量汇总统计；②对数据中的所有分类变量进行二维列联表分析；③以是否具有音乐特长为主分类变量，制作 3 个分类变量的三维列联表。

表 3.13　某艺术学校学生的特长情况

编号	性别	是否具有音乐特长	是否具有体育特长	是否具有美术特长
1	男	否	否	否
2	女	是	否	否
3	女	是	否	是
4	女	是	否	否
5	女	否	是	是
…	…			…
98	女	是	是	否
99	女	是	是	是
100	男	否	否	否

第 4 章　Stata 参数检验

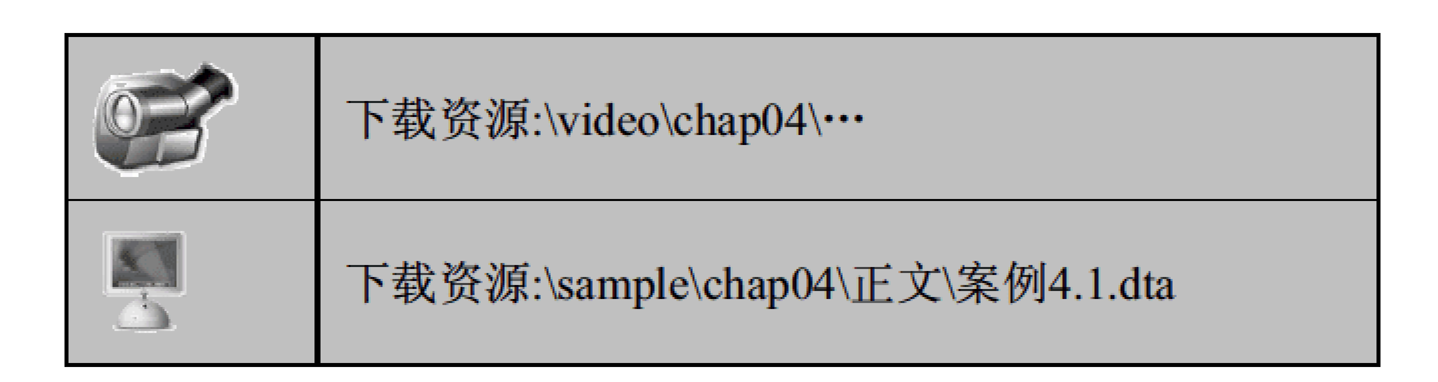
参数检验（Parameter Test）是指对参数的平均值、方差、比率等特征进行的统计检验。参数检验一般假设统计总体的具体分布为已知，但是其中的一些参数或者取值范围不确定，分析的主要目的是估计这些未知参数的取值，或者对这些参数进行假设检验。参数检验不仅能够对总体的特征参数进行推断，还能够对两个或多个总体的参数进行比较。常用的参数检验包括单一样本 T 检验、独立样本 T 检验、配对样本 T 检验、单一样本方差和双样本方差的假设检验等。下面我们通过实例的方式一一介绍这几种方法在 Stata 14.0 中的具体操作。

4.1　实例一——单一样本T检验

4.1.1　单一样本 T 检验的功能与意义

单一样本 T 检验（One-Samples T Test）是假设检验中最基本也是最常用的方法之一。与所有的假设检验一样，其依据的基本原理也是统计学中的“小概率反证法”原理。通过单一样本 T 检验，我们可以实现样本均值和总体均值的比较。检验的基本程序是首先提出原假设和备择假设，规定好检验的显著性水平，然后确定适当的检验统计量，并计算检验统计量的值，最后依据计算值和临界值的比较结果做出统计决策。

4.1.2　相关数据来源



	下载资源:\video\chap04\…
	下载资源:\sample\chap04\正文\案例4.1.dta

【例 4.1】河南省某高校 5 年前对大四学生体检时，发现学生的平均体重是 67.4kg。最近又抽查测量了该校 53 名大四学生的体重，如表 4.1 所示。试用 Stata 14.0 的单一样本 T 检验操作命令判断该校大四学生的体重与 5 年前相比是否有显著差异（设定显著性水平为 5%）。

表 4.1　河南省某高校 53 名大四学生的体重表

编号	体重/kg
001	62.7
002	57.3
003	52.6
004	61.8
005	60.8

（续表）

编号	体重/kg
…	…
051	51.2
052	63.6
053	64.5

4.1.3　Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有一个变量：体重。我们把体重变量设定为 weight，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作我们在第 1 章中已有详细讲述。录入完成后数据如图 4.1 所示。

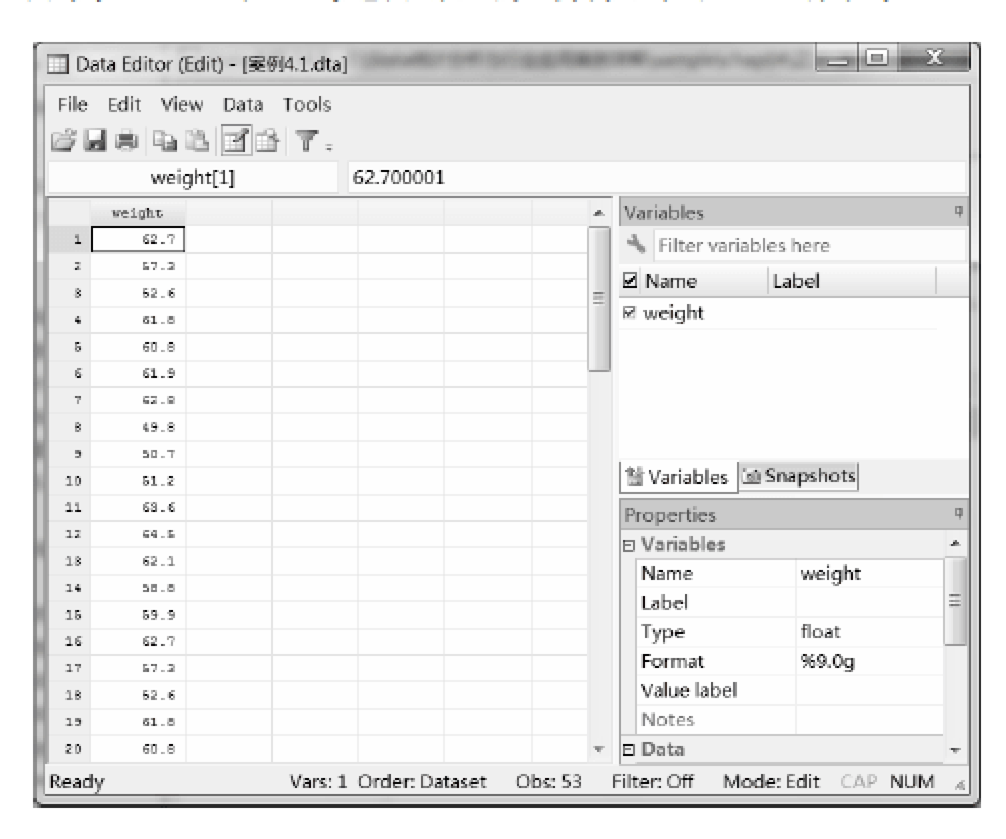


图 4.1　案例 4.1 数据

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

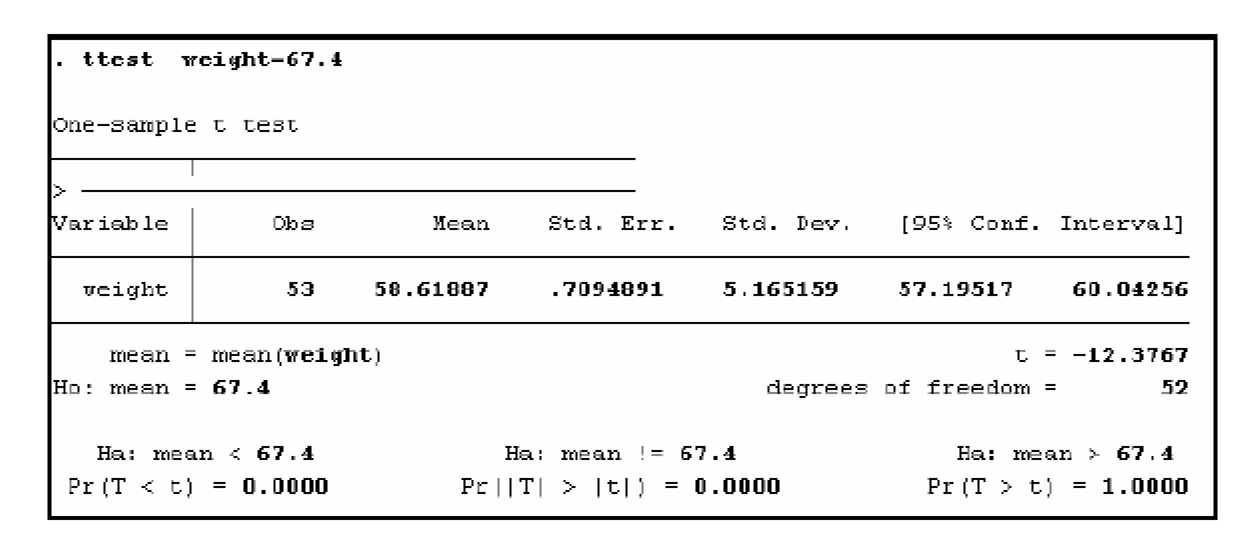
02 在主界面的“Command”文本框中输入命令：

```
ttest  weight=67.4
```

03 设置完毕后，按键盘上的回车键，等待输出结果。

4.1.4　结果分析

在 Stata 14.0 主界面的结果窗口我们可以看到如图 4.2 所示的分析结果。



```
. ttest weight=67.4

One-sample t test

Variable |  Obs     Mean    Std. Err.  Std. Dev.   [95% Conf. Interval]

  weight |   53   58.61887   .7094091   5.165159    57.19517    60.04256

    mean = mean(weight)                                   t = -12.3767
Ho: mean = 67.4                          degrees of freedom =       52

   Ha: mean < 67.4          Ha: mean != 67.4            Ha: mean > 67.4
 Pr(T < t) = 0.0000    Pr(|T| > |t|) = 0.0000        Pr(T > t) = 1.0000
```

图 4.2　分析结果图

通过观察分析结果，我们可以看出共有 53 个有效样本参与了假设检验，样本的均值是58.61887，标准差是 5.165159，方差的标准误是 0.7094891，95%的置信区间是[57.19517, 60.04256]，样本的 t 值为-12.3767，自由度为 52，Pr(|T| > |t|) = 0.0000，远小于 0.05，需要拒绝原假设，也就是说，该校大四学生的体重与 5 年前相比有显著差异。

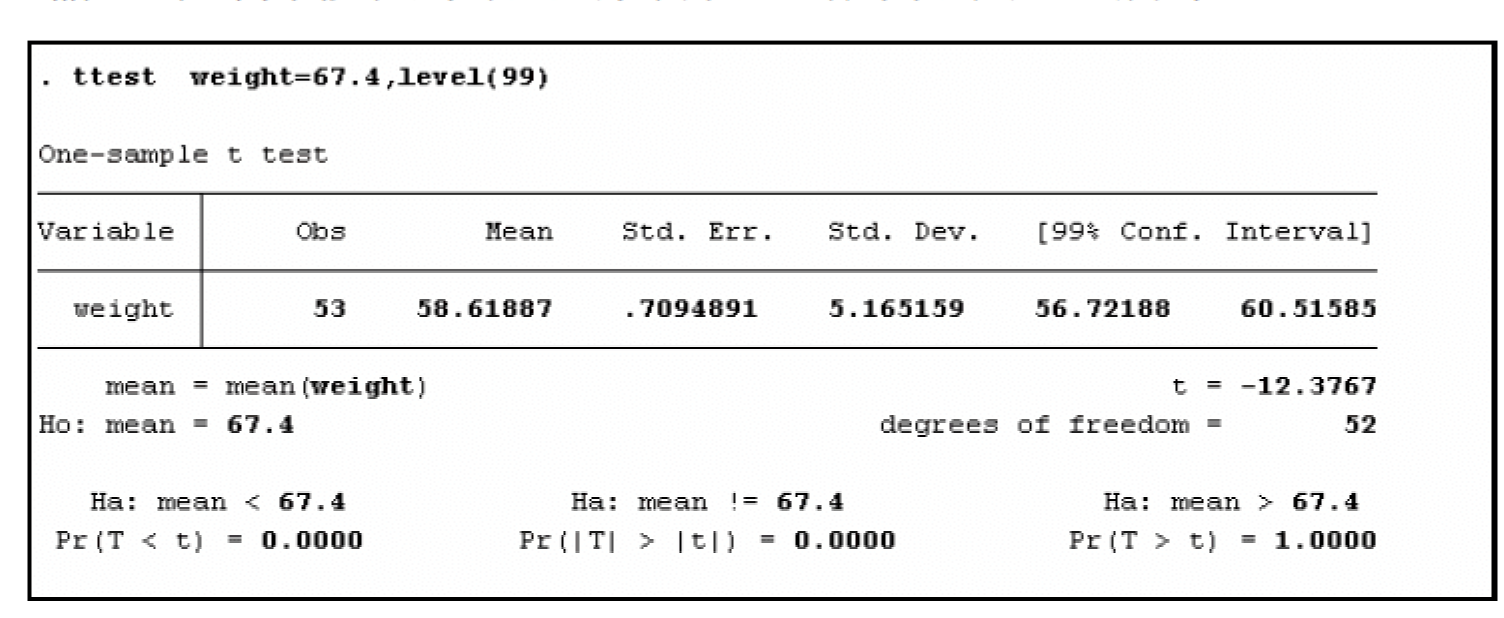
4.1.5 案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

例如，我们要把显著性水平调到 1%，也就是说置信水平为 99%，那么操作命令可以相应地修改为：

```
ttest  weight=67.4,level(99)
```

在命令窗口输入命令并按回车键进行确认，结果如图 4.3 所示。



```
. ttest  weight=67.4,level(99)

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [99% Conf. Interval]
---------+--------------------------------------------------------------------
  weight |      53    58.61887    .7094891    5.165159    56.72188    60.51585
------------------------------------------------------------------------------
    mean = mean(weight)                                           t = -12.3767
Ho: mean = 67.4                                  degrees of freedom =       52

   Ha: mean < 67.4             Ha: mean != 67.4               Ha: mean > 67.4
 Pr(T < t) = 0.0000       Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

图 4.3 分析结果图

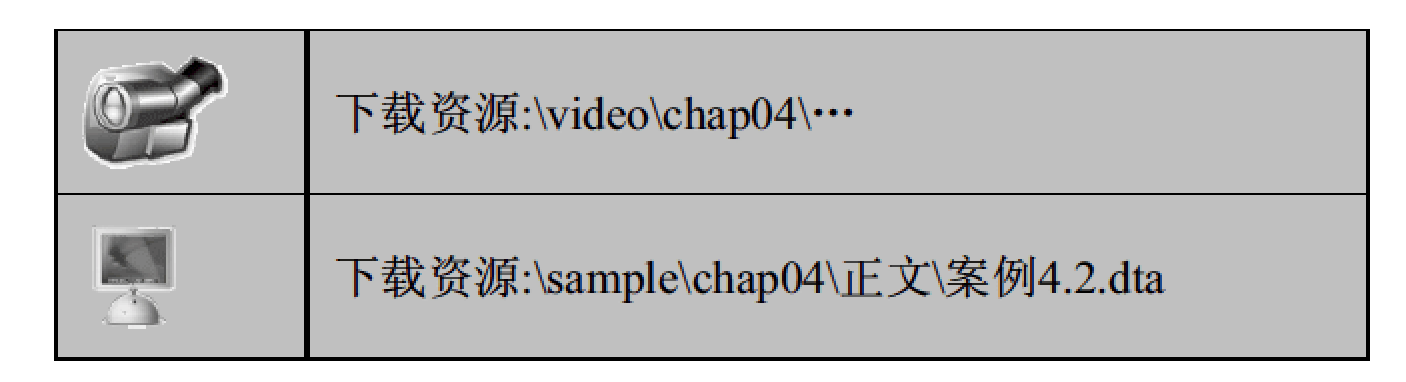
从上面的分析结果中可以看出与 95%的置信水平不同的地方在于置信区间得到了进一步的放大，这是正常的结果，因为这是要取得更高置信水平所必须付出的代价。

4.2 实例二——独立样本T检验

4.2.1 独立样本 T 检验的功能与意义

Stata 的独立样本 T 检验过程（Independent-Samples T Test）也是假设检验中最基本、最常用的方法之一。跟所有的假设检验一样，其依据的基本原理也是统计学中的“小概率反证法”原理。通过独立样本 T 检验，我们可以实现两个独立样本的均值比较。独立样本 T 检验过程的基本程序也是首先提出原假设和备择假设，规定好检验的显著性水平，然后确定适当的检验统计量，并计算检验统计量的值，最后依据计算值和临界值的比较结果做出统计决策。

4.2.2　相关数据来源



下载资源:\video\chap04\…

下载资源:\sample\chap04\正文\案例4.2.dta

【例 4.2】表 4.2 给出了 A、B 两所学校各 40 名高三学生的高考英语成绩。试用独立样本 T 检验方法研究两所学校被调查的高三学生的高考英语成绩之间有无明显的差别（设定显著性水平为 5%）。

表 4.2　A、B 两所学校各 40 名高三学生的高考英语成绩

编号	学校	高考英语成绩
001	A	145
002	A	147
003	A	139
004	A	138
005	A	135
…	…	…
078	B	105
079	B	99
080	B	108

4.2.3　Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有两个变量，分别是 A 学校高考英语成绩和 B 学校高考英语成绩。我们把 A 学校高考英语成绩变量设定为 englishA，把 B 学校高考英语成绩变量设定为 englishB，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作我们在第 1 章中已有详细讲述。录入完成后数据如图 4.4 所示。

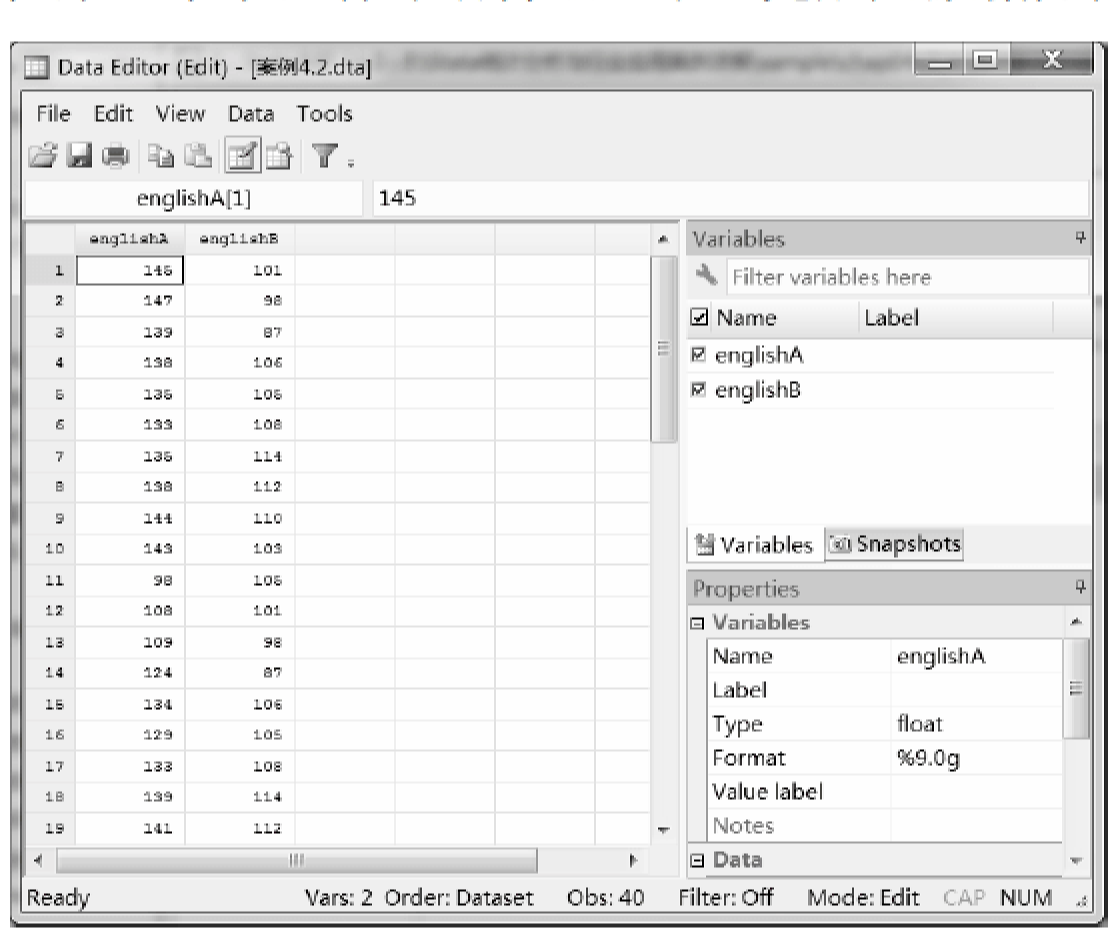


图 4.4　案例 4.2 数据

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入操作命令，并按键盘上的回车键进行确认。本例中对应的命令如下：

```
ttest englishA = englishB, unpaired
```

4.2.4 结果分析

在 Stata 14.0 主界面的结果窗口我们可以看到如图 4.5 所示的分析结果。

```
. ttest  englishA=englishB,unpaired

Two-sample t test with equal variances

Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+---------------------------------------------------------------------
englishA |      40     135.175    1.850463    11.70336    131.4321    138.9179
englishB |      40      104.95     1.09717    6.939112    102.7308    107.1692
---------+---------------------------------------------------------------------
combined |      80    120.0625    2.008317    17.96293     116.065      124.06
---------+---------------------------------------------------------------------
    diff |             30.225    2.151278                25.94213    34.50787
---------------------------------------------------------------------------------
    diff = mean(englishA) - mean(englishB)                        t =  14.0498
Ho: diff = 0                                     degrees of freedom =       78

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

图 4.5　分析结果图

通过观察分析结果，我们可以看出共有 80 个有效样本参与了假设检验，自由度为 78，其中变量 englishA 包括 40 个样本，均值为 135.175，标准差为 11.70336，标准误为 1.850463，95%的置信区间是[131.4321,138.9179]；变量 englishB 包括 40 个样本，均值为 104.95，标准差为 6.939112，标准误为 1.09717，95%的置信区间是[102.7308,107.1692]。Pr(|T| > |t|) = 0.0000 远小于 0.05，需要拒绝原假设，也就是说，两所学校被调查的高三学生的高考英语成绩之间存在明显的差别。

4.2.5 案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

1．延伸 1：改变置信水平

与单一样本 T 检验类似，例如我们要把显著性水平调到 1%，也就是说置信水平为 99%，那么操作命令可以相应地修改为：

```
ttest  englishA=englishB,unpaired level(99)
```

在命令窗口输入命令并按回车键进行确认，结果如图 4.6 所示。

```
. ttest  englishA=englishB,unpaired level(99)

Two-sample t test with equal variances
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [99% Conf. Interval]
---------+--------------------------------------------------------------------
englishA |      40     135.175    1.850463    11.70336    130.1641    140.1859
englishB |      40      104.95     1.09717    6.939112     101.979     107.921
---------+--------------------------------------------------------------------
combined |      80    120.0625    2.008317    17.96293    114.7615    125.3635
---------+--------------------------------------------------------------------
    diff |             30.225    2.151278                24.54489    35.90511
------------------------------------------------------------------------------
    diff = mean(englishA) - mean(englishB)                        t =  14.0498
Ho: diff = 0                                     degrees of freedom =       78

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

图 4.6　分析结果图

从上面的分析结果中可以看出与 95%的置信水平不同的地方在于置信区间得到了进一步的放大，这是正常的结果，因为这是要取得更高置信水平所必须付出的代价。

2．延伸 2：在异方差假定条件下进行假设检验

上面的检验过程是假定两个样本代表的总体之间存在相同的方差，如果假定两个样本代表的总体之间的方差并不相同，那么操作命令可以相应地修改为：

```
ttest  englishA=englishB,unpaired level(99) unequal
```

在命令窗口输入命令并按回车键进行确认，结果如图 4.7 所示。

```
. ttest  englishA=englishB,unpaired level(99) unequal

Two-sample t test with unequal variances
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [99% Conf. Interval]
---------+--------------------------------------------------------------------
englishA |      40     135.175    1.850463    11.70336    130.1641    140.1859
englishB |      40      104.95     1.09717    6.939112     101.979     107.921
---------+--------------------------------------------------------------------
combined |      80    120.0625    2.008317    17.96293    114.7615    125.3635
---------+--------------------------------------------------------------------
    diff |             30.225    2.151278                24.51203    35.93797
------------------------------------------------------------------------------
    diff = mean(englishA) - mean(englishB)                        t =  14.0498
Ho: diff = 0                     Satterthwaite's degrees of freedom =  63.4048

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

图 4.7　分析结果图

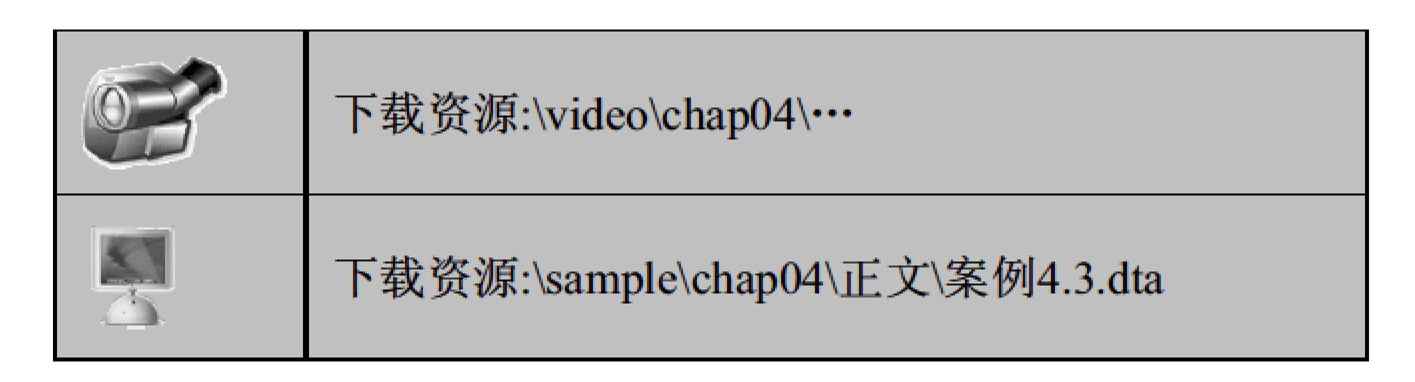
可以看出在本例中同方差假定和异方差假定之间的结果没有差别。

4.3　实例三——配对样本T检验

4.3.1　配对样本 T 检验的功能与意义

Stata 的配对样本 T 检验过程（Paired-Samples T Test）也是假设检验中的方法之一。与所有的假设检验一样，其依据的基本原理也是统计学中的“小概率反证法”原理。通过配对样本 T 检验，我们可以实现对成对数据的样本均值比较。其与独立样本 T 检验的区别是：两个样本来自于同一总体，而且数据的顺序不能调换。配对样本 T 检验过程的基本程序也是首先提出原假设和备择假设，规定好检验的显著性水平，然后确定适当的检验统计量，并计算检验统计量的值，最后依据计算值和临界值的比较结果做出统计决策。

4.3.2 相关数据来源



【例 4.3】为了研究一种减肥药品的效果，特抽取了 30 名试验者进行试验，服用该产品一个疗程前后的体重如表 4.3 所示。试用配对样本 T 检验的方法判断该药物能否引起试验者体重的明显变化（设定显著性水平为 5%）。

表 4.3 试验者服药前后的体重（单位：kg）

编号	服药前体重	服药后体重
001	88.6	75.6
002	85.2	76.5
003	75.2	68.2
004	78.4	67.2
005	76	69.9
…	……	…
048	82.7	78.1
049	82.4	75.3
050	75.6	69.9

4.3.3 Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有两个变量，分别是服药前体重和服药后体重。我们把服药前体重变量设定为 qian，把服药后体重变量设定为 hou，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作我们在第 1 章中已有详细讲述。录入完成后数据如图 4.8 所示。

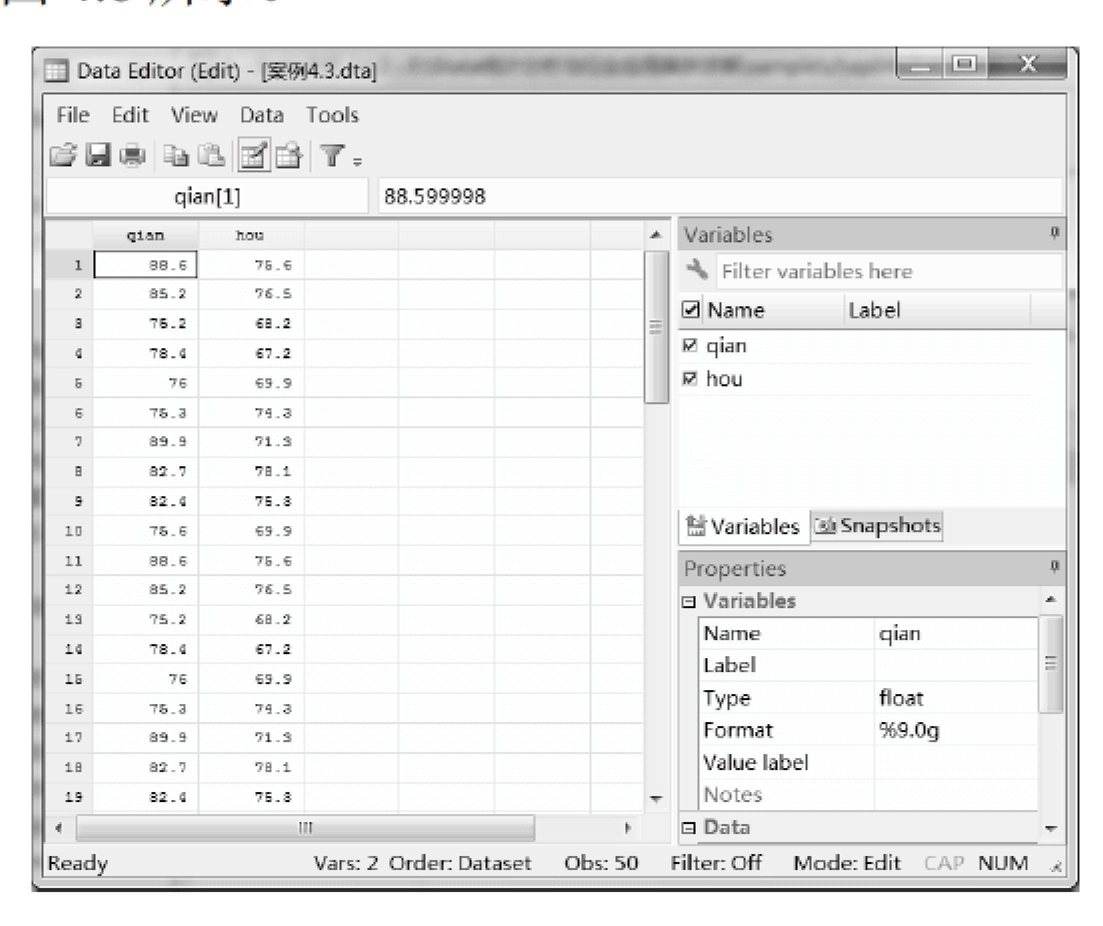


图 4.8 案例 4.3 数据

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入命令：

```
ttest  qian=hou
```

03 设置完毕后，按键盘上的回车键，等待输出结果。

4.3.4 结果分析

在 Stata 14.0 主界面的结果窗口我们可以看到如图 4.9 所示的分析结果。

```
. ttest  qian=hou

Paired t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
    qian |      50       80.93    .7646007    5.406543    79.39348    82.46652
     hou |      50       72.63    .5139305    3.634037    71.59722    73.66278
---------+--------------------------------------------------------------------
    diff |      50    8.299999    .6677101    4.721423    6.958186    9.641813
------------------------------------------------------------------------------
     mean(diff) = mean(qian - hou)                                t =  12.4305
 Ho: mean(diff) = 0                              degrees of freedom =       49

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

图 4.9 分析结果图

通过观察分析结果，我们可以看出共有 50 对有效样本参与了假设检验，自由度为 48，其中变量 qian 包括 50 个样本，均值为 80.93，标准差为 5.406543，标准误为 0.7646007，95%的置信区间是[79.39348,82.46652]；变量 hou 包括 50 个样本，均值为 72.63，标准差为 3.634037，标准误为 0.5139305，95%的置信区间是[71.59722,73.66278]。Pr(|T| > |t|) = 0.0000，远小于 0.05，所以需要拒绝原假设，也就是说，该药物能引起试验者体重的明显变化。

4.3.5 案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

与单一样本 T 检验类似，例如我们要把显著性水平调到 1%，也就是说置信水平为 99%，那么操作命令可以相应地修改为：

```
ttest  qian=hou,level(99)
```

在命令窗口输入命令并按回车键进行确认，结果如图 4.10 所示。

```
. ttest  qian=hou,level(99)

Paired t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [99% Conf. Interval]
---------+--------------------------------------------------------------------
    qian |      50       80.93    .7646007    5.406543    78.88091    82.97909
     hou |      50       72.63    .5139305    3.634037    71.25269    74.00731
---------+--------------------------------------------------------------------
    diff |      50    8.299999    .6677101    4.721423    6.510568    10.08943
------------------------------------------------------------------------------
     mean(diff) = mean(qian - hou)                                t =  12.4305
 Ho: mean(diff) = 0                              degrees of freedom =       49

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

图 4.10 分析结果图

从上面的分析结果中可以看出与 95%的置信水平不同的地方在于置信区间得到了进一步的放大，这是正常的结果，因为这是要取得更高置信水平所必须付出的代价。

4.4 实例四——单一样本方差的假设检验

4.4.1 单一样本方差假设检验的功能与意义

方差的概念用来反映波动情况，常用于质量控制与市场波动等情形。单一总体方差的假设检验的基本程序也是首先提出原假设和备择假设，规定好检验的显著性水平，然后确定适当的检验统计量，并计算检验统计量的值，最后依据计算值和临界值的比较结果做出统计决策。

4.4.2 相关数据来源

	下载资源:\video\chap04\…
	下载资源:\sample\chap04\正文\案例4.4.dta

【例 4.4】为研究某只股票的收益率波动情况，某课题组对该只股票连续 60 天的收益率情况进行了调查研究，调查得到的数据经整理后如表 4.4 所示。试对该数据资料进行假设检验其方差是否等于 1（设定显著性水平为 5%）。

表 4.4　某只股票的收益率波动情况

编号	收益率
1	0.136 984
2	-0.643 22
3	0.557 802
4	-0.604 79
5	0.684 176
…	…
58	-0.171 8
59	0.290 384
60	-0.628 38

4.4.3 Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有一个变量：收益率。我们把收益率变量设定为 return，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作我们在第 1 章中已有详细讲述。录入完成后数据如图 4.11 所示。

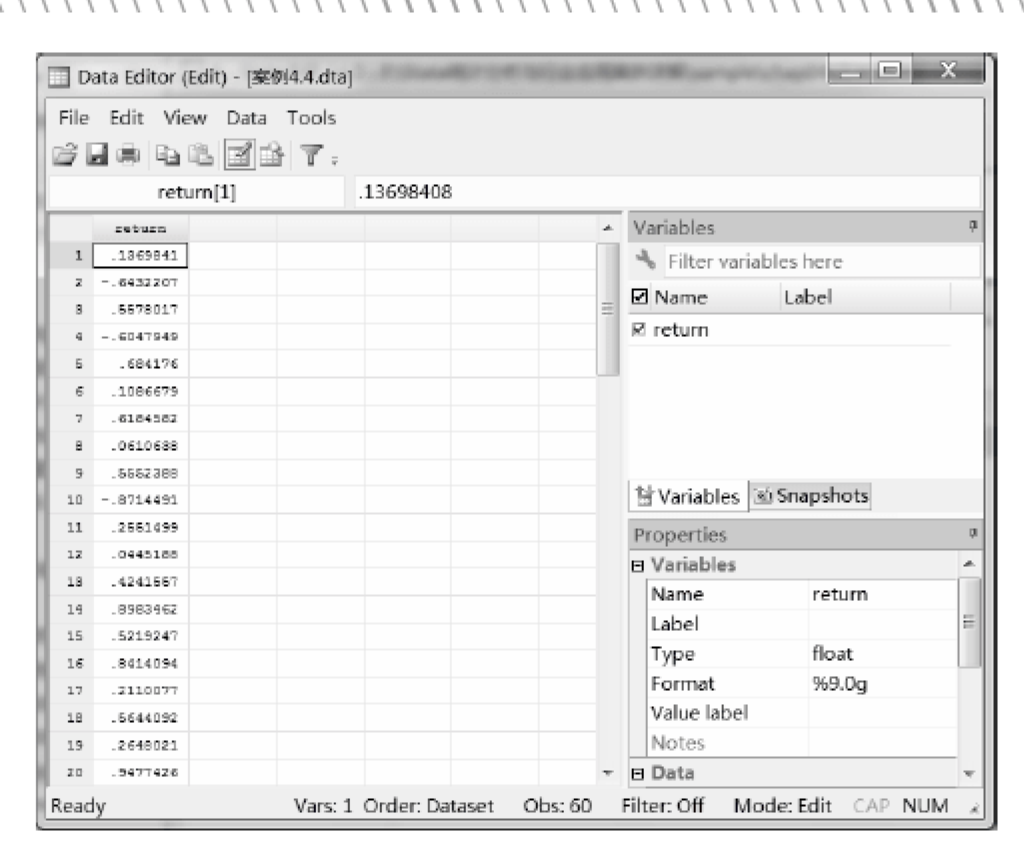


图 4.11　案例 4.4 数据

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

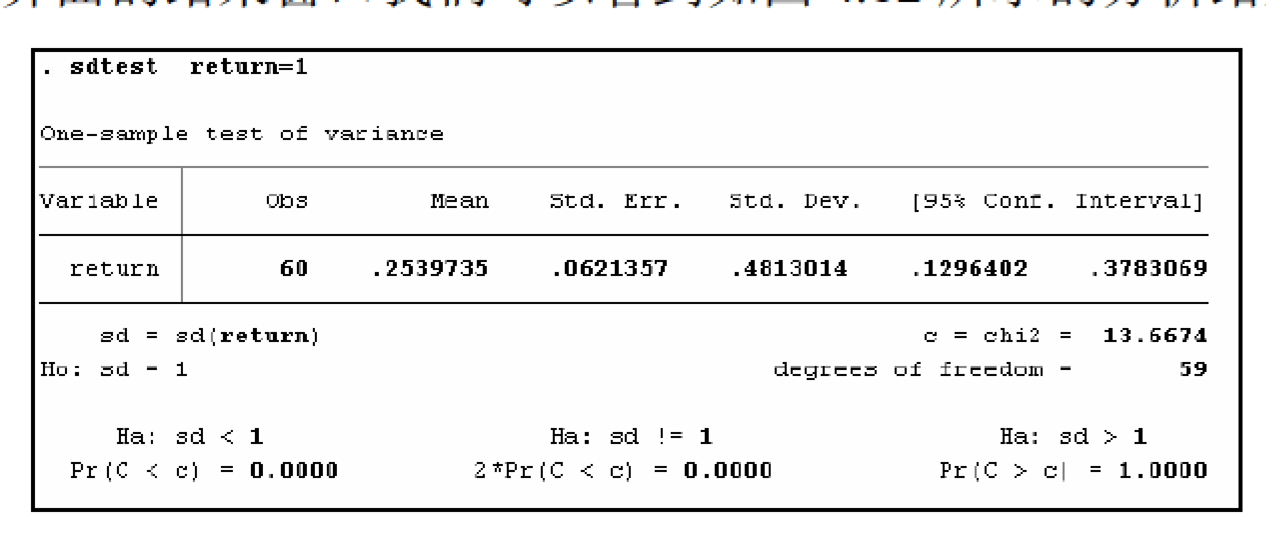
02 在主界面的“Command”文本框中输入命令：

```
sdtest  return=1
```

03 设置完毕后，按键盘上的回车键，等待输出结果。

4.4.4　结果分析

在 Stata 14.0 主界面的结果窗口我们可以看到如图 4.12 所示的分析结果。



```
. sdtest  return=1

One-sample test of variance

Variable |     Obs       Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+---------------------------------------------------------------------
  return |      60   .2539735    .0621357    .4813014    .1296402    .3783069
-------------------------------------------------------------------------------
    sd = sd(return)                                       c = chi2 =  13.5674
Ho: sd = 1                                        degrees of freedom =       59

    Ha: sd < 1                 Ha: sd != 1                   Ha: sd > 1
 Pr(C < c) = 0.0000      2*Pr(C < c) = 0.0000            Pr(C > c) = 1.0000
```

图 4.12　分析结果图

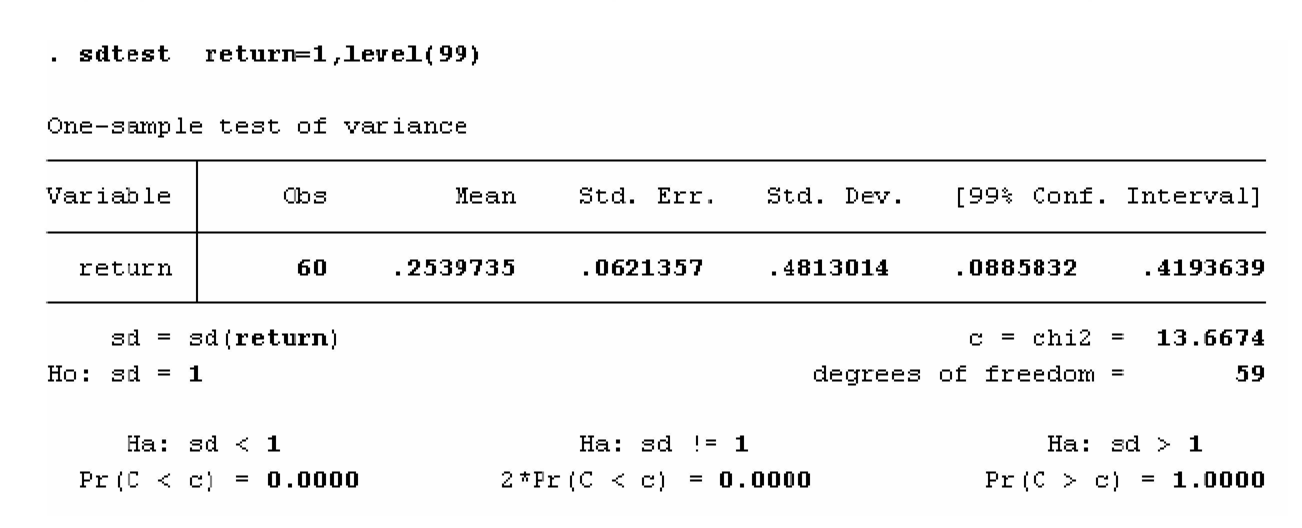
通过观察分析结果，我们可以看出共有 60 个有效样本参与了假设检验，自由度为 59，均值为 0.2539735，标准差为 0.4813014，标准误为 0.0621357，95%的置信区间是[0.1296402, 0.3783069]。2*Pr(C < c) = 0.0000，远小于 0.05，所以需要拒绝原假设，也就是说，该股票的收益率方差不显著等于 1。

4.4.5　案例延伸

例如，我们要把显著性水平调到 1%，也就是说置信水平为 99%，那么操作命令可以相应地修改为：

```
sdtest  return=1,level(99)
```

在命令窗口输入命令并按回车键进行确认，结果如图 4.13 所示。



```
. sdtest  return=1,level(99)

One-sample test of variance

Variable |   Obs      Mean    Std. Err.  Std. Dev.   [99% Conf. Interval]
---------+------------------------------------------------------------------
  return |    60   .2539735   .0621357   .4813014    .0885832    .4193639
---------+------------------------------------------------------------------
    sd = sd(return)                                     c = chi2 =  13.6674
Ho: sd = 1                                    degrees of freedom =       59

    Ha: sd < 1                 Ha: sd != 1                  Ha: sd > 1
  Pr(C < c) = 0.0000        2*Pr(C < c) = 0.0000         Pr(C > c) = 1.0000
```

图 4.13　分析结果图

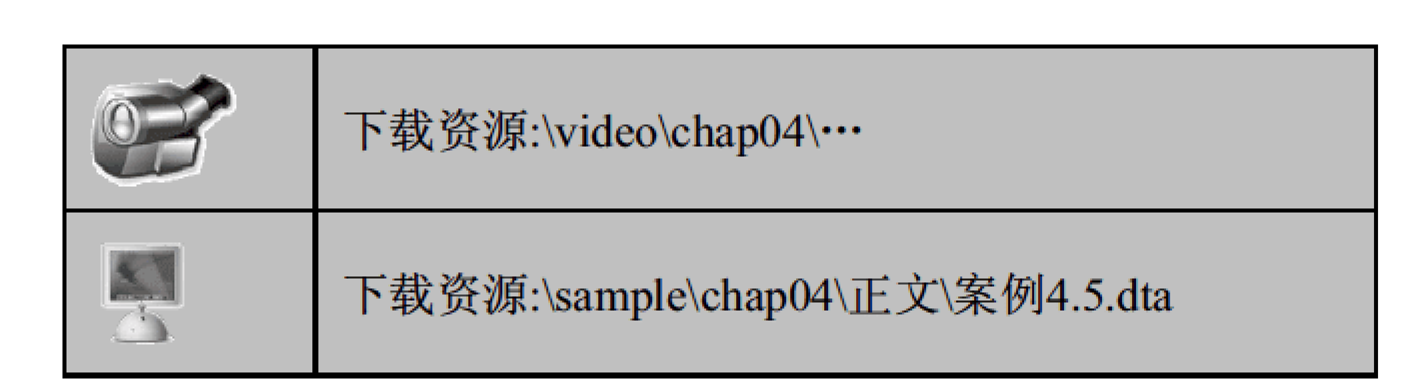
从上面的分析结果中可以看出与 95%的置信水平不同的地方在于：置信区间得到了进一步的放大，这是正常的结果，因为这是要取得更高置信水平所必须付出的代价。

4.5　实例五——双样本方差的假设检验

4.5.1　双样本方差假设检验的功能与意义

双样本方差假设检验用来判断两个样本的波动情况是否相同。它的基本程序也是首先提出原假设和备择假设，规定好检验的显著性水平，然后确定适当的检验统计量，并计算检验统计量的值，最后依据计算值和临界值的比较结果做出统计决策。

4.5.2　相关数据来源



	下载资源:\video\chap04\…
	下载资源:\sample\chap04\正文\案例4.5.dta

【例 4.5】为研究某两只股票的收益率波动情况是否相同，某课题组对这两只股票连续 30 天的收益率情况进行了调查研究，调查得到的数据经整理后如表 4.5 所示。试使用 Stata 14.0 对该数据资料进行假设，检验其方差是否相同（设定显著性水平为 5%）。

表 4.5　某两只股票的收益率波动情况

编号	收益率A	收益率B
1	0.136 984	0.715 281
2	0.643 221	0.699 069
3	0.557 802	0.232 269
4	0.604 795	0.098 188
5	0.684 176	0.594 84
…	…	…
28	0.894 475	0.171 803
29	0.058 066	0.290 384
30	0.675 949	0.628 377

4.5.3 Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有两个变量，分别为收益率 A 和收益率 B。我们把收益率 A 变量设定为 returnA，把收益率 B 变量设定为 returnB，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作我们在第 1 章中已有详细讲述。录入完成后数据如图 4.14 所示。

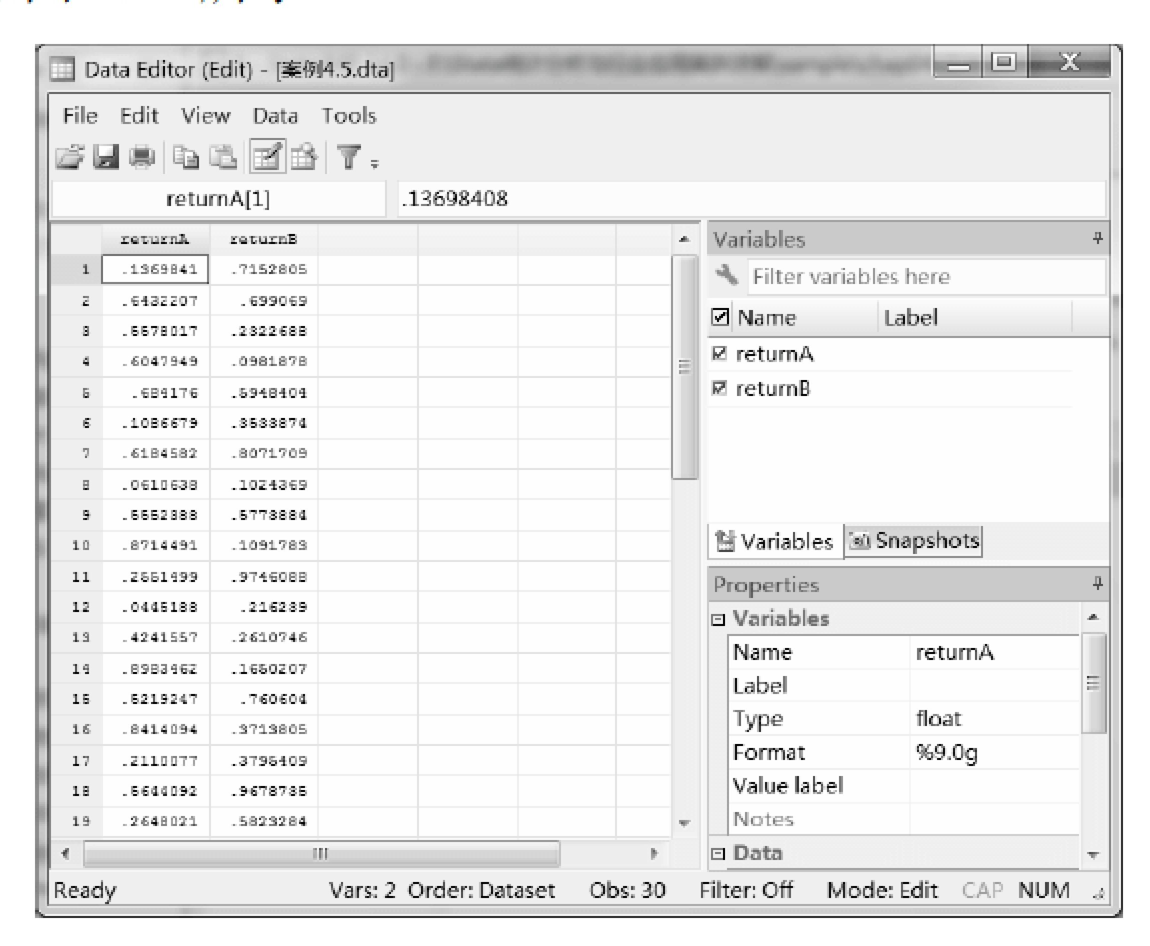


图 4.14　案例 4.5 数据

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入命令：

```
sdtest  returnA= returnB
```

03 设置完毕后，按键盘上的回车键，等待输出结果。

4.5.4 结果分析

在 Stata 14.0 主界面的结果窗口我们可以看到如图 4.15 所示的分析结果。

```
. sdtest  returnA= returnB

Variance ratio test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
 returnA |      30    .4907723    .0522183    .2860114    .3839739    .5975707
 returnB |      30    .4291026    .0526941    .2886173    .3213311    .5368741
---------+--------------------------------------------------------------------
combined |      60    .4599374    .0369953    .2865641    .3859101    .5339648
------------------------------------------------------------------------------
    ratio = sd(returnA) / sd(returnB)                             f =   0.9820
Ho: ratio = 1                                    degrees of freedom =   29, 29

    Ha: ratio < 1               Ha: ratio != 1                 Ha: ratio > 1
  Pr(F < f) = 0.4807         2*Pr(F < f) = 0.9614           Pr(F > f) = 0.5193
```

图 4.15　分析结果图

通过观察分析结果，我们可以看出共有 30 对有效样本参与了假设检验，自由度为 29，其中变量 returnA 包括 30 个样本，均值为 0.4907723，标准差为 0.2860114，标准误为 0.0522183，95%的置信区间是[0.3839739,0.5975707]；变量 returnB 包括 30 个样本，均值为 0.4291026，标准差为 0.2886173，标准误为 0.0526941，95%的置信区间是[0.3213311,0.5368741]。2*Pr(F<f) = 0.9614，远大于 0.05，所以需要接受原假设，也就是说，两只股票的收益率波动情况显著相同。

4.5.5 案例延伸

例如，我们要把显著性水平调到 1%，也就是说置信水平为 99%，那么操作命令可以相应地修改为：

```
sdtest  returnA= returnB,level(99)
```

在命令窗口输入命令并按回车键进行确认，结果如图 4.16 所示。

```
. sdtest  returnA= returnB,level(99)

Variance ratio test

Variable |     Obs       Mean    Std. Err.   Std. Dev.   [99% Conf. Interval]
---------+--------------------------------------------------------------------
 returnA |      30   .4907723    .0522183    .2860114    .3468385     .634706
 returnB |      30   .4291026    .0526941    .2886173    .2838574    .5743478
---------+--------------------------------------------------------------------
combined |      60   .4599374    .0369953    .2865641     .361465    .5584099
------------------------------------------------------------------------------
    ratio = sd(returnA) / sd(returnB)                             f =   0.9820
Ho: ratio = 1                                    degrees of freedom =   29, 29

    Ha: ratio < 1               Ha: ratio != 1                 Ha: ratio > 1
  Pr(F < f) = 0.4807         2*Pr(F < f) = 0.9614           Pr(F > f) = 0.5193
```

图 4.16　分析结果图

从上面的分析结果中可以看出与 95%的置信水平不同的地方在于置信区间得到了进一步的放大，这是正常的结果，因为这是要取得更高置信水平所必须付出的代价。

4.6　本章习题

（1）江西省某高校 3 年前对大二学生体检时，发现学生的平均身高是 175 厘米。最近又抽查测量了该校 63 名大二学生的身高，如表 4.6 所示。试用 Stata 14.0 的单一样本 T 检验操作命令判断该校大二学生的身高与 3 年前相比是否有显著差异（设定显著性水平为 5%）。

表 4.6　江西省某高校 63 名大二学生的身高数据

编号	身高（cm）
001	164.5
002	162.1
003	158.8
004	159.9
005	162.7
…	…
061	151.2
062	163.6
063	164.5

（2）表 4.7 给出了 X、Y 两所学校各 38 名初三学生的中考语文成绩。试用独立样本 T 检验方法研究两所学校被调查的初三学生的中考语文成绩之间有无明显的差别（设定显著性水平为 5%）。

表 4.7　X、Y 两所学校各 38 名初三学生的中考语文成绩

编号	学校	中考语文成绩
001	X	103
002	X	105
003	X	101
004	X	98
005	X	87
…	…	…
074	Y	135
075	Y	138
076	Y	144

（3）为了研究一种杀虫剂的效果，特抽取了 30 平方米的麦田进行试验，其使用该产品前后的含虫量如表 4.8 所示。试用配对样本 T 检验的方法判断该杀虫剂是否有效（设定显著性水平为 5%）。

表 4.8　使用杀虫剂前后的含虫量（单位：个/平方米）

编号	使用杀虫剂前	使用杀虫剂后
001	18	12
002	20	8
003	15	7
004	16	15
005	12	18
…	…	…
028	11	11
029	10	10
030	10	10

（4）为研究某基金的收益率波动情况，某课题组对该基金连续 50 天的收益率情况进行了调查研究，调查得到的数据经整理后如表 4.9 所示。试对该数据资料进行假设，检验其方差是否等于 1（设定显著性水平为 5%）。

表 4.9　某基金的收益率波动情况

编号	收益率
1	0.564 409
2	0.264 802
3	0.947 743
4	0.276 915
5	0.118 016
…	…
48	-0.967 87
49	0.582 328
50	0.795 3

（5）为研究某两只基金的收益率波动情况是否相同，某课题组对这两只基金连续 20 天的收益率情况进行了调查研究，调查得到的数据经整理后如表 4.10 所示。试使用 Stata 14.0 对该数据资料进行假设，检验其方差是否相同（设定显著性水平为 5%）。

表 4.10 某两只基金的收益率波动情况

编号	收益率A	收益率B
1	0.424 156	0.261 075
2	0.898 346	0.165 021
3	0.521 925	0.760 604
4	0.841 409	0.371 381
5	0.211 008	0.379 541
…	…	…
18	0.564 409	0.967 874
19	0.264 802	0.582 328
20	0.947 743	0.7953

第 5 章　Stata 非参数检验

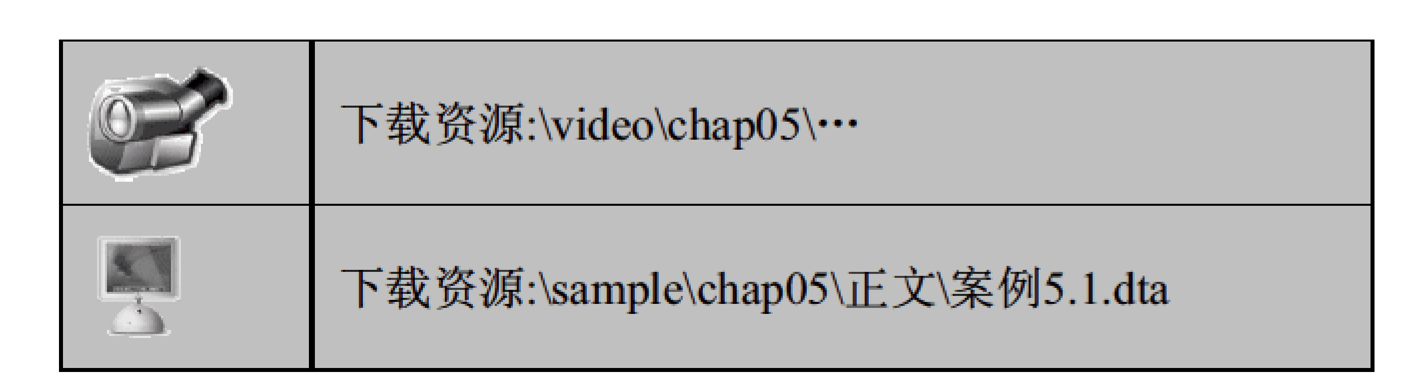
一般情况下，参数检验方法假设统计总体的具体分布为已知，但是我们往往会遇到一些总体分布不能用有限个实参数来描述或者不考虑被研究的对象为何种分布，以及无法合理假设总体分布形式的情形，这时我们就需要放弃对总体分布参数的依赖，从而去寻求更多来自样本的信息，基于这种思路的统计检验方法被称为非参数检验。常用的非参数检验（Nonparametric Tests）包括单样本正态分布检验、两独立样本检验、两相关样本检验、多独立样本检验、游程检验等。下面我们将一一介绍这些方法在实例中的应用。

5.1　实例一——单样本正态分布检验

5.1.1　单样本正态分布检验的功能与意义

单样本正态分布检验本质上属于一种拟合优度检验，基本功能是通过检验样本特征来探索总体是否服从正态分布。Stata 的单样本正态分布检验有很多种，常用的包括偏度-峰度检验、Wilks-Shapiro 两种。

5.1.2　相关数据来源



	下载资源:\video\chap05\…
	下载资源:\sample\chap05\正文\案例5.1.dta

【例 5.1】表 5.1 给出了山东财经大学某专业 60 名男生的百米速度。试用单样本正态分布检验方法研究其是否服从正态分布。

表 5.1　百米速度

编号	速度/m/s
001	15.1
002	15.2
003	12.4
004	12.4
005	12.6
…	…
058	12.6
059	12.6
060	13.7

5.1.3 Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有一个变量，即速度。我们把速度变量设定为 speed，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作我们在第 1 章中已有详细讲述。录入完成后数据如图 5.1 所示。

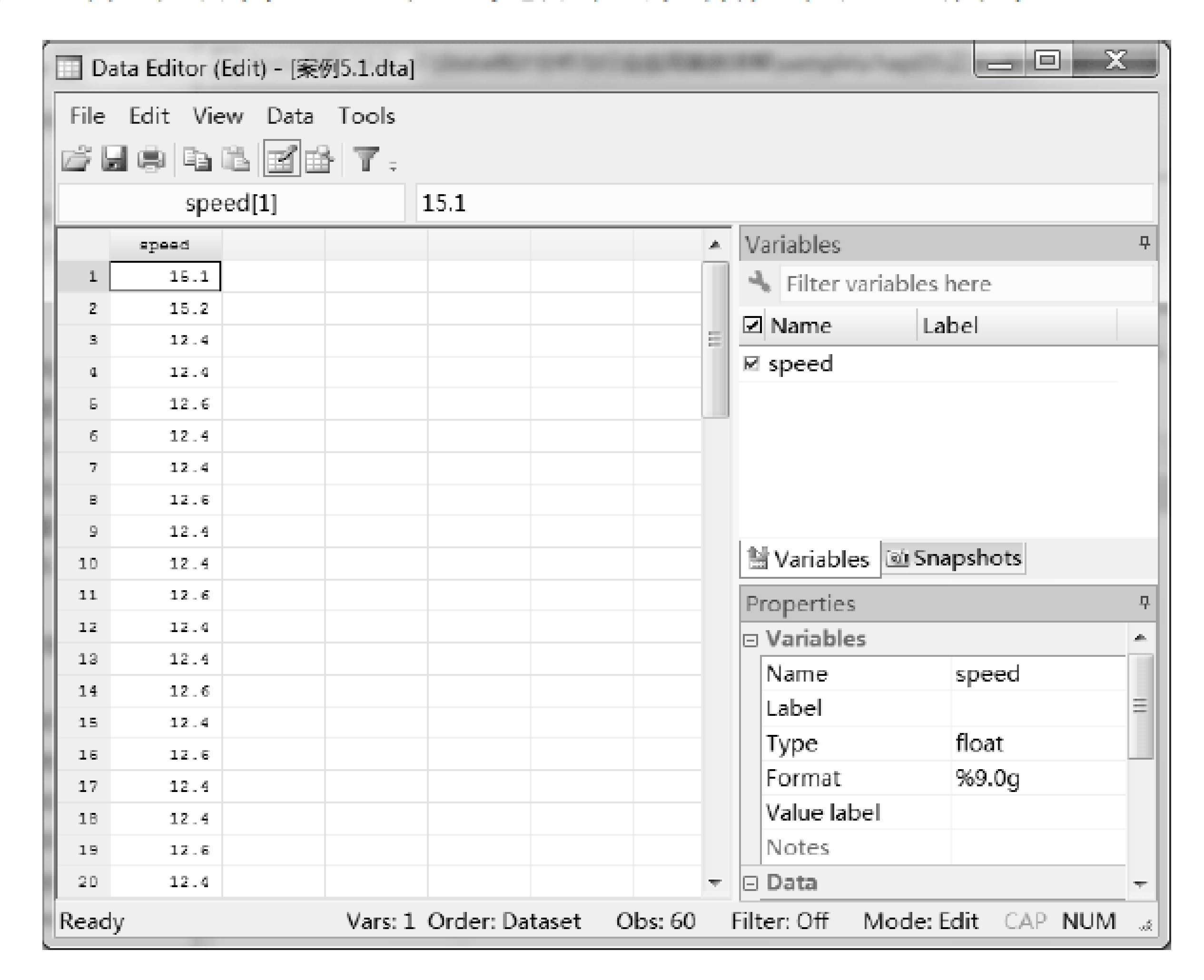


图 5.1　案例 5.1 数据

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

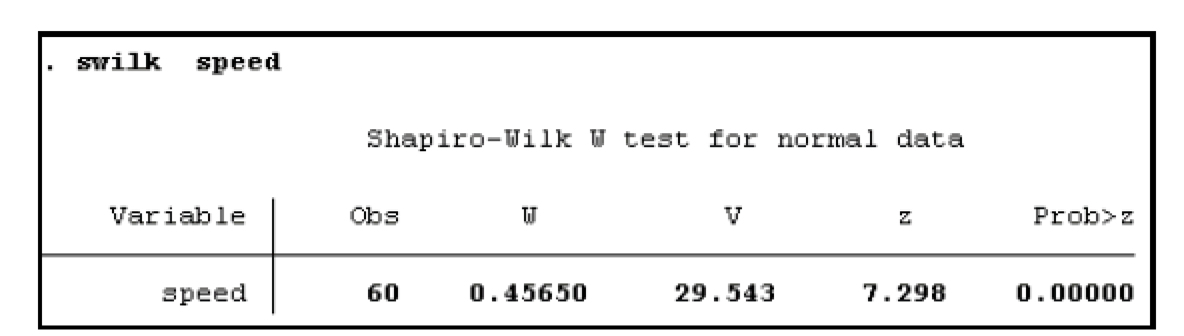
02 Wilks-Shapiro、偏度-峰度检验两种检验方式在主界面的“Command”文本框中输入的命令格式分别如下。

- swilk speed：本命令的含义是对 speed 变量使用 Wilks-Shapiro 检验方式进行单样本正态分布检验。
- sktest speed：本命令的含义是对 speed 变量使用偏度-峰度检验方式进行单样本正态分布检验。

03 设置完毕后，按键盘上的回车键，等待输出结果。

5.1.4 结果分析

在 Stata 14.0 主界面的结果窗口我们可以看到如图 5.2 和图 5.3 所示的分析结果。

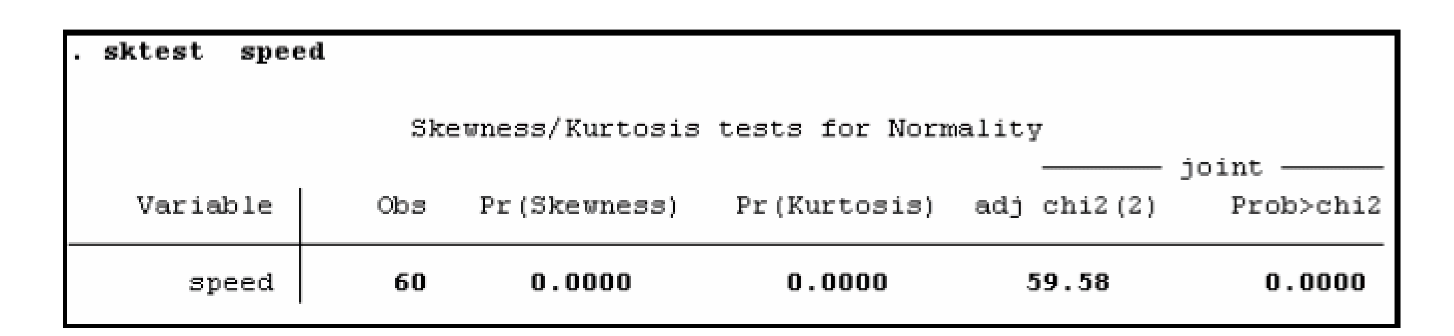


```
. swilk speed

                   Shapiro-Wilk W test for normal data
```

Variable	Obs	W	V	z	Prob>z
speed	60	0.45650	29.543	7.298	0.00000

图 5.2　分析结果图



```
. sktest  speed

                    Skewness/Kurtosis tests for Normality
                                                         ------- joint ------
    Variable |       Obs  Pr(Skewness)  Pr(Kurtosis) adj chi2(2)  Prob>chi2
-------------+---------------------------------------------------------------
       speed |        60     0.0000        0.0000        59.58       0.0000
```

图 5.3　分析结果图

通过观察分析结果，我们可以看出两种检验方法的检验结果是一致的，共有 60 个有效样本参与了假设检验，P 值均远小于 0.05，所以需要拒绝原假设，也就是说，百米速度数据不服从正态分布。

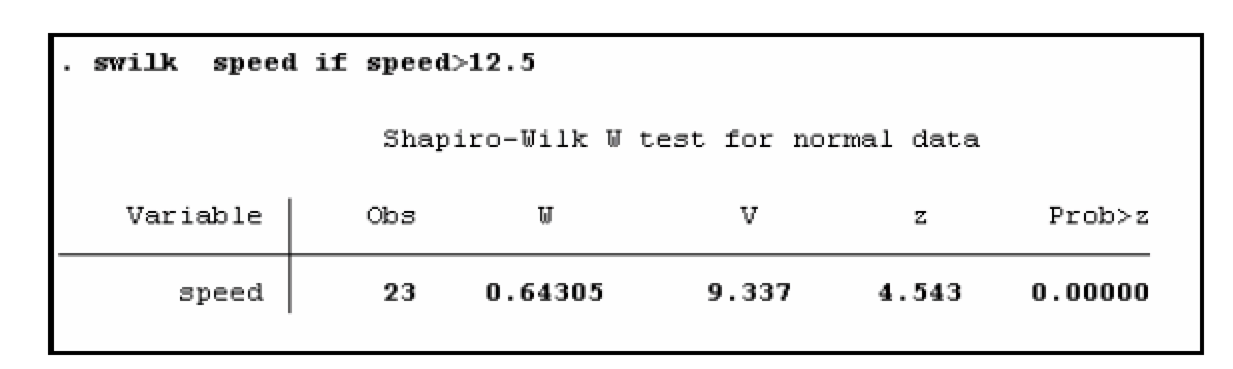
5.1.5　案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

例如，我们只针对 speed 变量大于 12.5 的观测样本进行单样本正态分布检验，那么操作命令即为：

```
swilk  speed if speed>12.5
```

在命令窗口输入命令并按回车键进行确认，结果如图 5.4 所示。



```
. swilk  speed if speed>12.5

                   Shapiro-Wilk W test for normal data

    Variable |        Obs       W           V         z       Prob>z
-------------+------------------------------------------------------
       speed |         23    0.64305      9.337     4.543    0.00000
```

图 5.4　分析结果图

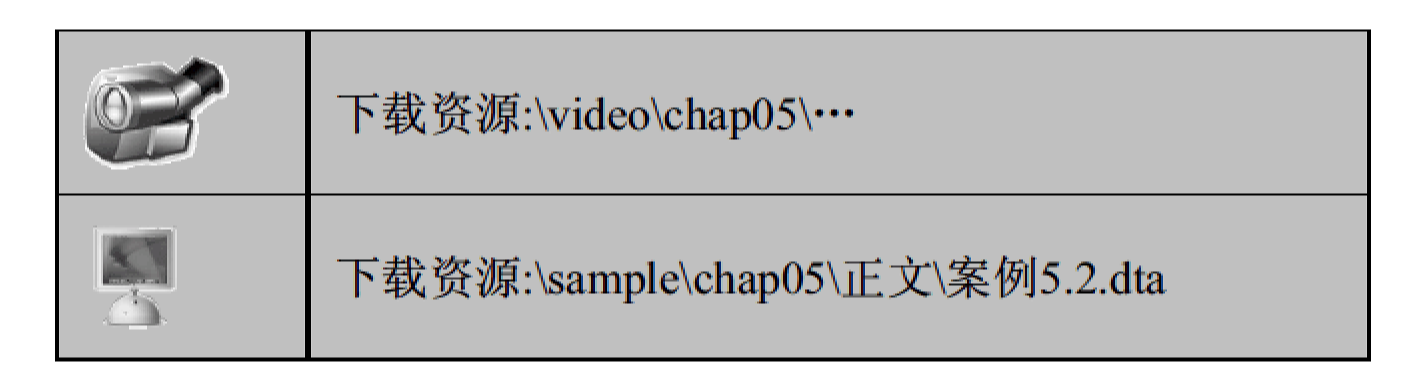
通过观察分析结果，我们可以看出共有 23 个有效样本参与了假设检验，P 值均远小于 0.05，所以需要拒绝原假设，也就是说，百米速度数据不服从正态分布。

5.2　实例二——两独立样本检验

5.2.1　两独立样本检验的功能与意义

跟前面的检验方法一样，Stata 的两独立样本检验（Two-Independent samples Test）也是非参数检验方法的一种，其基本功能是可以判断两个独立样本是否来自相同分布的总体。这种检验过程是通过分析两个独立样本的均数、中位数、离散趋势、偏度等描述性统计量之间的差异来实现的。

5.2.2 相关数据来源



【例 5.2】表 5.2 给出了广东省东北部和西北部主要年份的年降雨量。试用两独立样本检验方法判断两个地区的年降雨量是否存在显著差异。

表 5.2　广东省东北部和西北部主要年份年降雨量（单位：mm）

年份	降雨量	
	粤东北	粤西北
1980	1461.7	1586.1
1985	1607.8	1726.9
1990	1709.0	1284.8
1995	1171.0	1766.4
1996	1361.5	1693.1
1997	1847.5	1815.3
1998	1458.2	1737.5
1999	1033.8	1318.7
2000	1850.9	1318.2
2001	1560.3	1889.2
2002	1110.3	1480.9
2003	1415.2	1251.8

5.2.3 Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有 3 个变量，分别是年份、地区和降雨量。我们把年份变量设定为 year，把地区变量设定为 group 并且把粤东北定义为 1，把粤西北定义为 2，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作我们在第 1 章中已有详细讲述。录入完成后数据如图 5.5 所示。

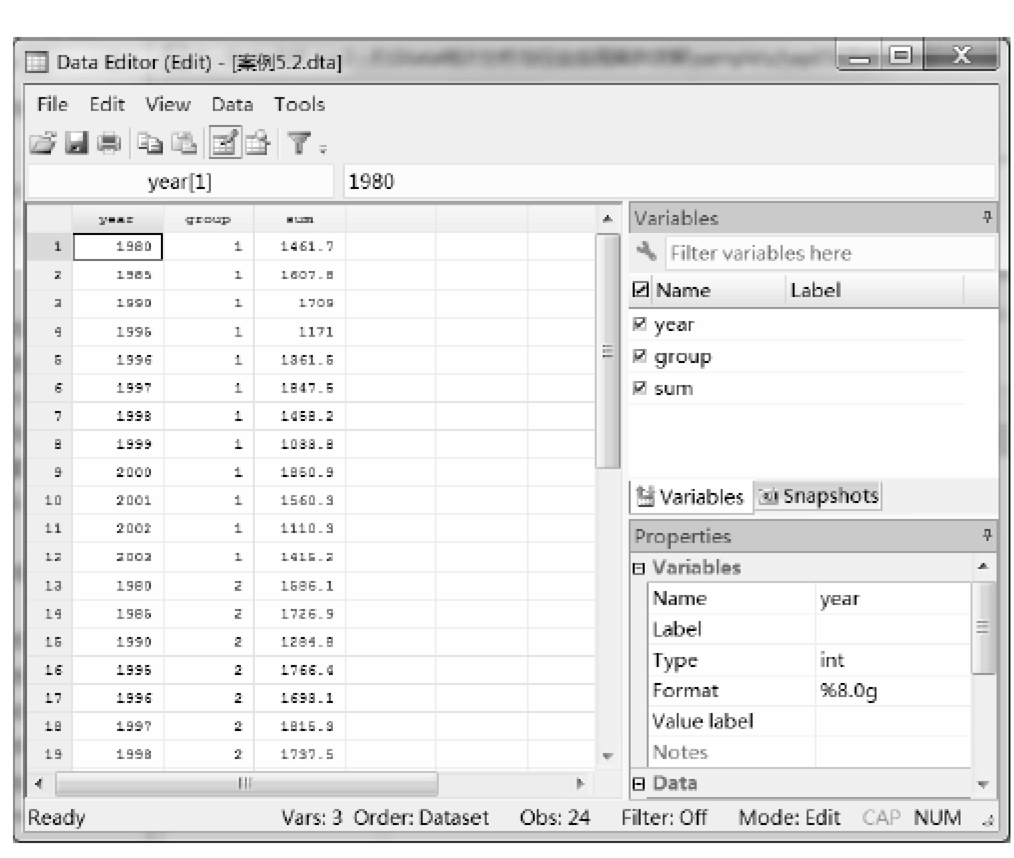


图 5.5　案例 5.2 数据

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入如下命令（旨在用两独立样本检验方法判断两个地区的年降雨量是否存在显著差异）：

```
ranksum  sum,by( group)
```

03 设置完毕后，按键盘上的回车键，等待输出结果。

5.2.4　结果分析

在 Stata 14.0 主界面的结果窗口我们可以看到如图 5.6 所示的分析结果。

通过观察分析结果，我们可以看出共有 24 个有效样本参与了假设检验，Prob > |z| = 0.3556，远大于 0.05，所以需要接受原假设，也就是说，两个地区的年降雨量存在显著差异。

5.2.5　案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

例如，我们只针对 year 变量大于 1990 的观测样本进行两独立样本检验，那么操作命令即为：

```
ranksum  sum if year>1990,by( group)
```

在命令窗口输入命令并按回车键进行确认，结果如图 5.7 所示。

```
. ranksum  sum,by( group)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

        group |      obs    rank sum    expected
-------------+---------------------------------
            1 |       12         134         150
            2 |       12         166         150
-------------+---------------------------------
     combined |       24         300         300

unadjusted variance      300.00
adjustment for ties        0.00
                     ----------
adjusted variance        300.00

Ho: sum(group==1) = sum(group==2)
             z =  -0.924
    Prob > |z| =   0.3556
```

图 5.6　分析结果图

```
. ranksum  sum if year>1990,by( group)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

        group |      obs    rank sum    expected
-------------+---------------------------------
            1 |        9          74        85.5
            2 |        9          97        85.5
-------------+---------------------------------
     combined |       18         171         171

unadjusted variance      128.25
adjustment for ties        0.00
                     ----------
adjusted variance        128.25

Ho: sum(group==1) = sum(group==2)
             z =  -1.015
    Prob > |z| =   0.3099
```

图 5.7　分析结果图

通过观察分析结果，我们可以看出共有 18 个有效样本参与了假设检验，Prob > |z| = 0.3099，远大于 0.05，所以需要接受原假设，也就是说，两个地区的年降雨量存在显著差异。

5.3 实例三——两相关样本检验

5.3.1 两相关样本检验的功能与意义

两相关样本检验（2-Related samples Test）的基本功能是可以判断两个相关的样本是否来自相同分布的总体。

5.3.2 相关数据来源

	下载资源:\video\chap05\…
	下载资源:\sample\chap05\正文\案例5.3.dta

【例 5.3】为分析一种新药的效果，特选取了 52 名病人进行试验，表 5.3 给出了试验者服药前后的血红蛋白数量。试用两相关样本检验方法判断该药能否引起患者体内血红蛋白数量的显著变化。

表 5.3 患者服药前后血红蛋白的数量变化

患者编号	服药前血红蛋白数量/g/L	服药后血红蛋白数量/g/L
001	13	12.5
002	12.6	11.4
003	13.1	12.5
004	12.9	13.9
005	11.5	11
…	…	…
050	13.4	14.1
051	15.2	13.6
052	10.9	11.5

5.3.3 Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有两个变量，分别是服药前血红蛋白数量和服药后血红蛋白数量。我们把服药前血红蛋白数量这一变量设定为 qian，把服药后血红蛋白数量这一变量设定为 hou，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作我们在第 1 章中已有详细讲述。录入完成后数据如图 5.8 所示。

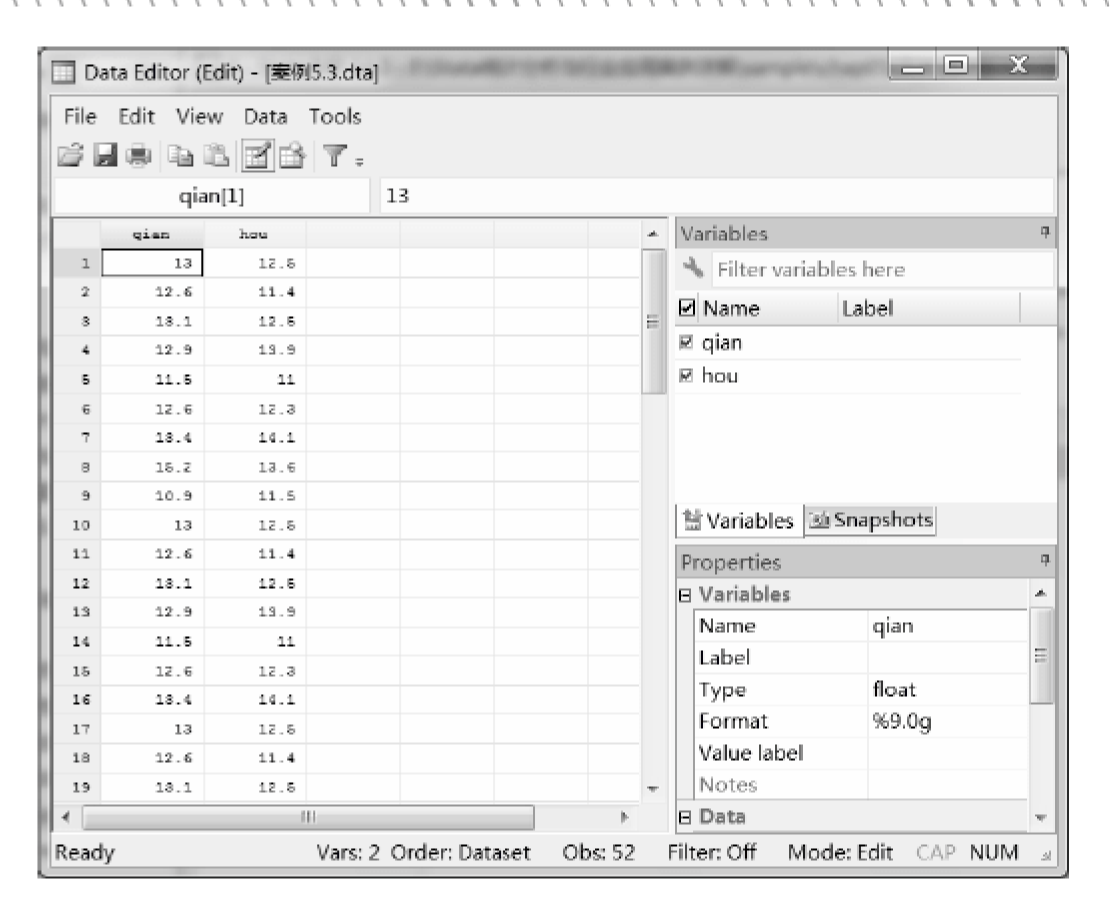


图 5.8　案例 5.3 数据

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

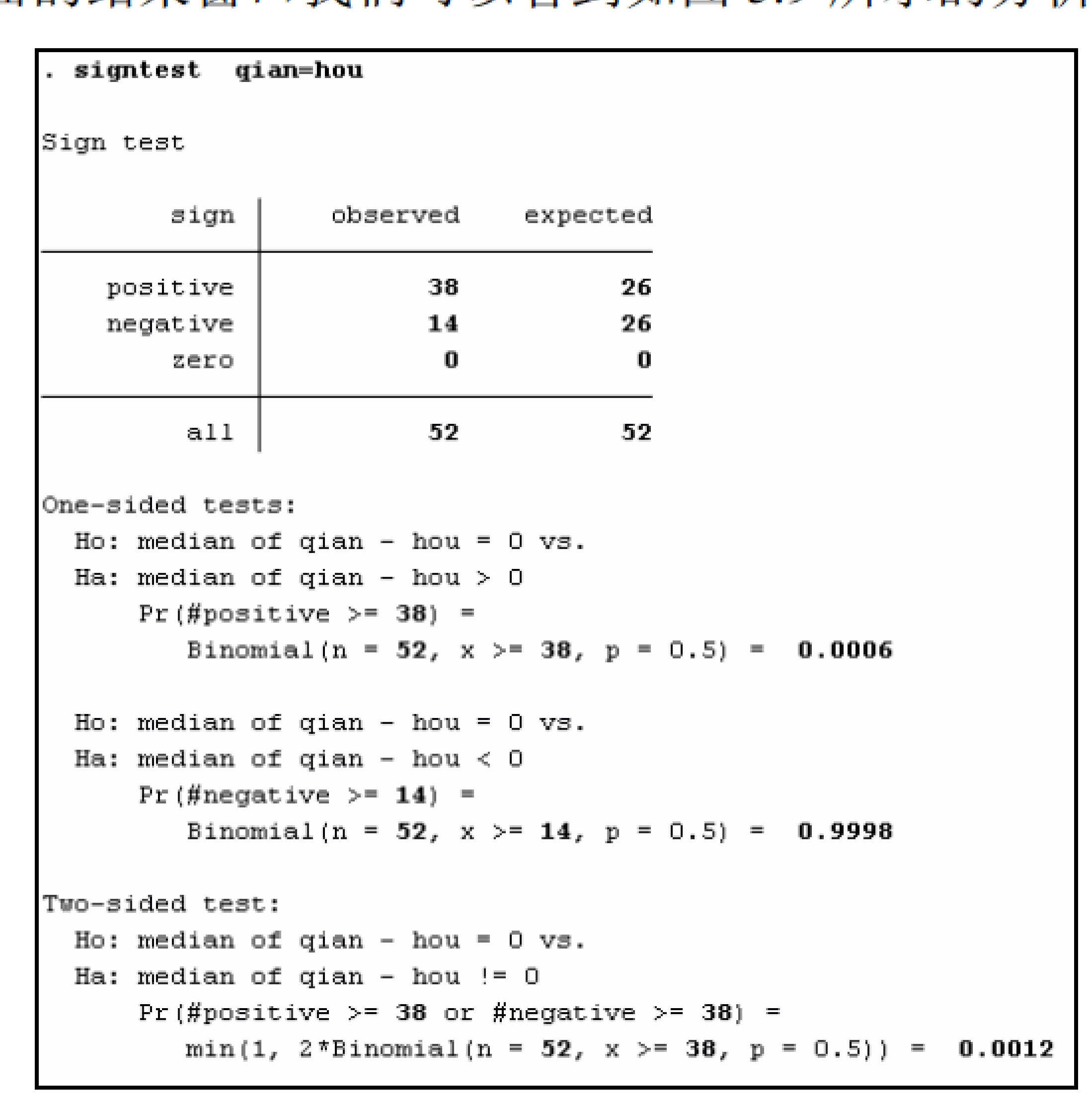
02 在主界面的“Command”文本框中输入如下命令（旨在使用两相关样本检验方法判断患者体内血红蛋白数量是否发生显著变化）：

```
signtest  qian=hou
```

03 设置完毕后，按键盘上的回车键，等待输出结果。

5.3.4　结果分析

在 Stata 14.0 主界面的结果窗口我们可以看到如图 5.9 所示的分析结果。



```
. signtest  qian=hou

Sign test

        sign |    observed    expected
-------------+------------------------
    positive |          38          26
    negative |          14          26
        zero |           0           0
-------------+------------------------
         all |          52          52

One-sided tests:
  Ho: median of qian - hou = 0 vs.
  Ha: median of qian - hou > 0
      Pr(#positive >= 38) =
         Binomial(n = 52, x >= 38, p = 0.5) =  0.0006

  Ho: median of qian - hou = 0 vs.
  Ha: median of qian - hou < 0
      Pr(#negative >= 14) =
         Binomial(n = 52, x >= 14, p = 0.5) =  0.9998

Two-sided test:
  Ho: median of qian - hou = 0 vs.
  Ha: median of qian - hou != 0
      Pr(#positive >= 38 or #negative >= 38) =
         min(1, 2*Binomial(n = 52, x >= 38, p = 0.5)) =  0.0012
```

图 5.9　分析结果图

可以看出本结论与通过检验均值得出的结论是一致的。本检验结果包括符号检验、单侧

检验和双侧检验 3 部分。符号检验（Sign test）的原理是通过用配对的两组数据做差，原假设是两组数据不存在显著差别，所以两组数据做差的结果应该是正数、负数大体相当。在本例中，期望值是有 26 个正数，26 个负数，然而实际的观察值却是 38 个正数，所以两组数据存在显著差异。也就是说该药引起了患者体内血红蛋白数量的显著变化。单侧检验和双侧检验的结果解读在前面章节多有涉及，这里不再赘述。

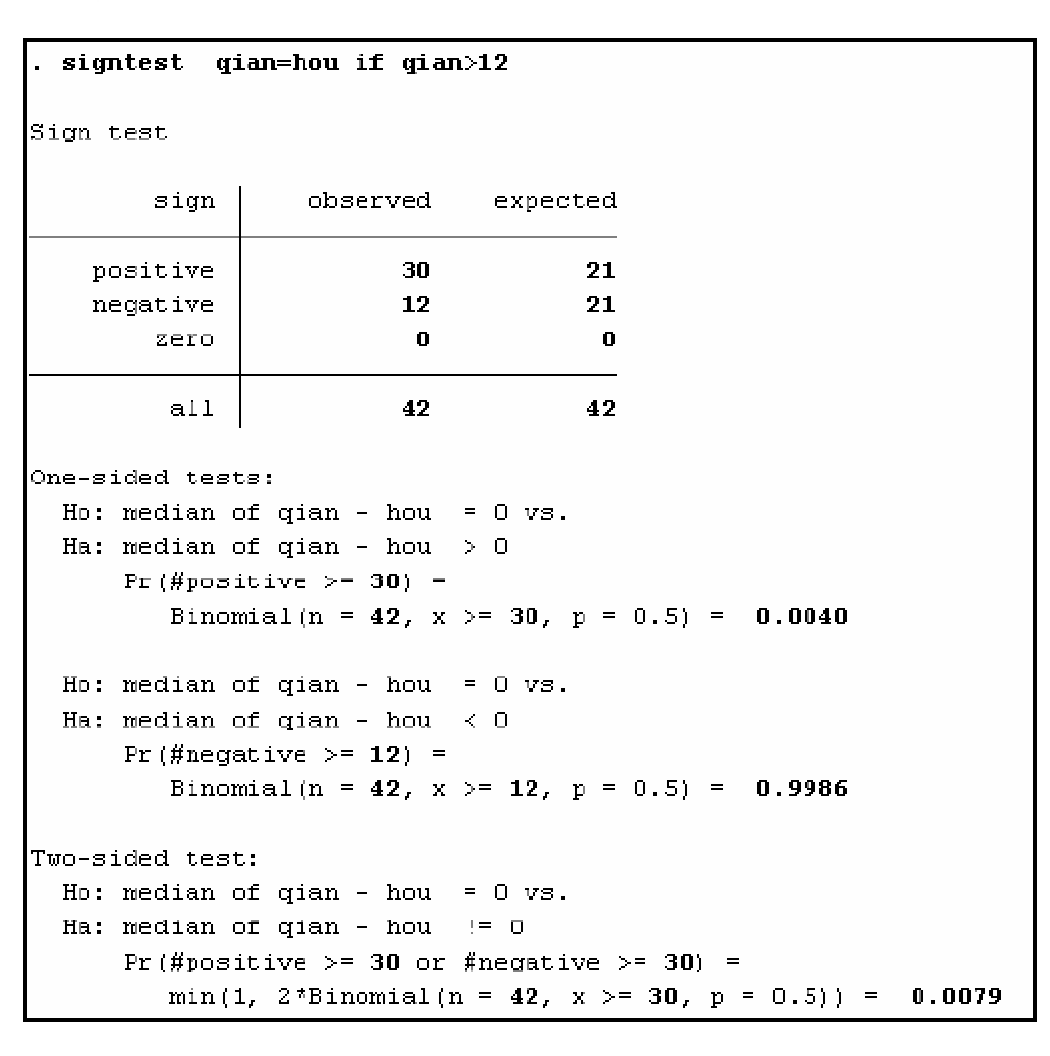
5.3.5 案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

例如，我们只针对 qian 变量大于 12 的观测样本进行两相关样本检验，那么操作命令即为：

```
signtest  qian=hou if qian>12
```

在命令窗口输入命令并按回车键进行确认，结果如图 5.10 所示。



```
. signtest  qian=hou if qian>12

Sign test

        sign |    observed    expected
-------------+------------------------
    positive |          30          21
    negative |          12          21
        zero |           0           0
-------------+------------------------
         all |          42          42

One-sided tests:
  Ho: median of qian - hou  = 0 vs.
  Ha: median of qian - hou  > 0
      Pr(#positive >= 30) =
         Binomial(n = 42, x >= 30, p = 0.5) =  0.0040

  Ho: median of qian - hou  = 0 vs.
  Ha: median of qian - hou  < 0
      Pr(#negative >= 12) =
         Binomial(n = 42, x >= 12, p = 0.5) =  0.9986

Two-sided test:
  Ho: median of qian - hou  = 0 vs.
  Ha: median of qian - hou  != 0
      Pr(#positive >= 30 or #negative >= 30) =
         min(1, 2*Binomial(n = 42, x >= 30, p = 0.5)) =  0.0079
```

图 5.10 分析结果图

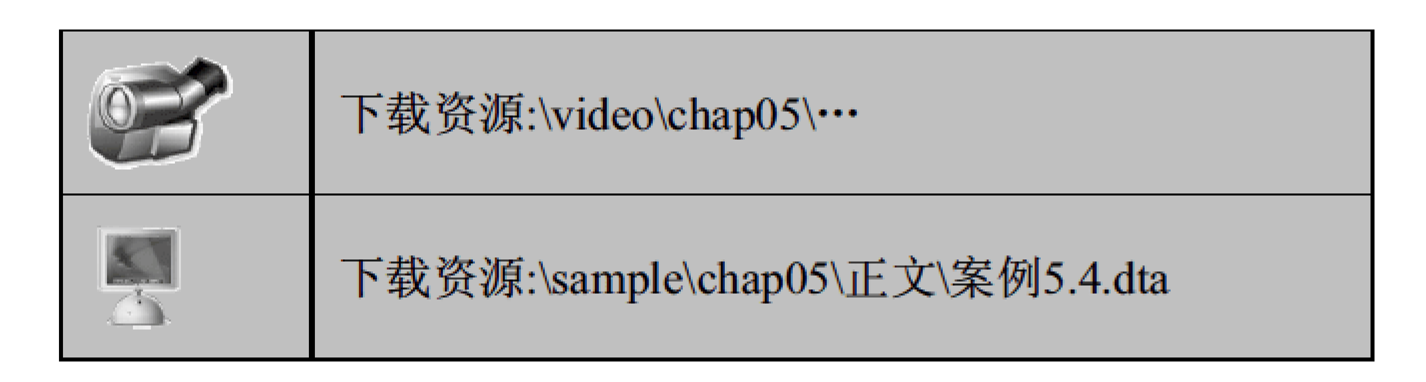
通过观察分析结果，我们可以看出期望值是有 21 个正数、21 个负数，然而实际的观察值却是 30 个正数，所以两组数据存在显著差异，也就是说该药引起了患者体内血红蛋白数量的显著变化。

5.4 实例四——多独立样本检验

5.4.1 多独立样本检验的功能与意义

顾名思义，多独立样本检验（K-Independent samples Test）用于判断多个独立的样本是否来自相同分布的总体。

5.4.2　相关数据来源



【例 5.4】某公司新招聘的一批员工毕业于 4 所不同的高校，并且来源于 4 所不同高校的员工构成了 4 个独立的样本。待到实习期结束后，高管对这些新员工进行考察打分，结果如表 5.4 所示。试用多独立样本检验方法分析毕业于不同高校的员工在工作上的表现是否有显著的差异。

表 5.4　员工考核成绩

A高校	89	97	84	86	…	90	89
B高校	75	76	73	71	…	70	71
C高校	59	52	54	51	…	53	55
D高校	32	29	28	25	…	18	31

5.4.3　Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有两个变量，分别为高校和分数。我们把分数变量设定为 goal，把高校变量设定为 school，并且把 A、B、C、D 共 4 所高校分别定义为 1、2、3、4，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作我们在第 1 章中已有详细讲述。录入完成后数据如图 5.11 所示。

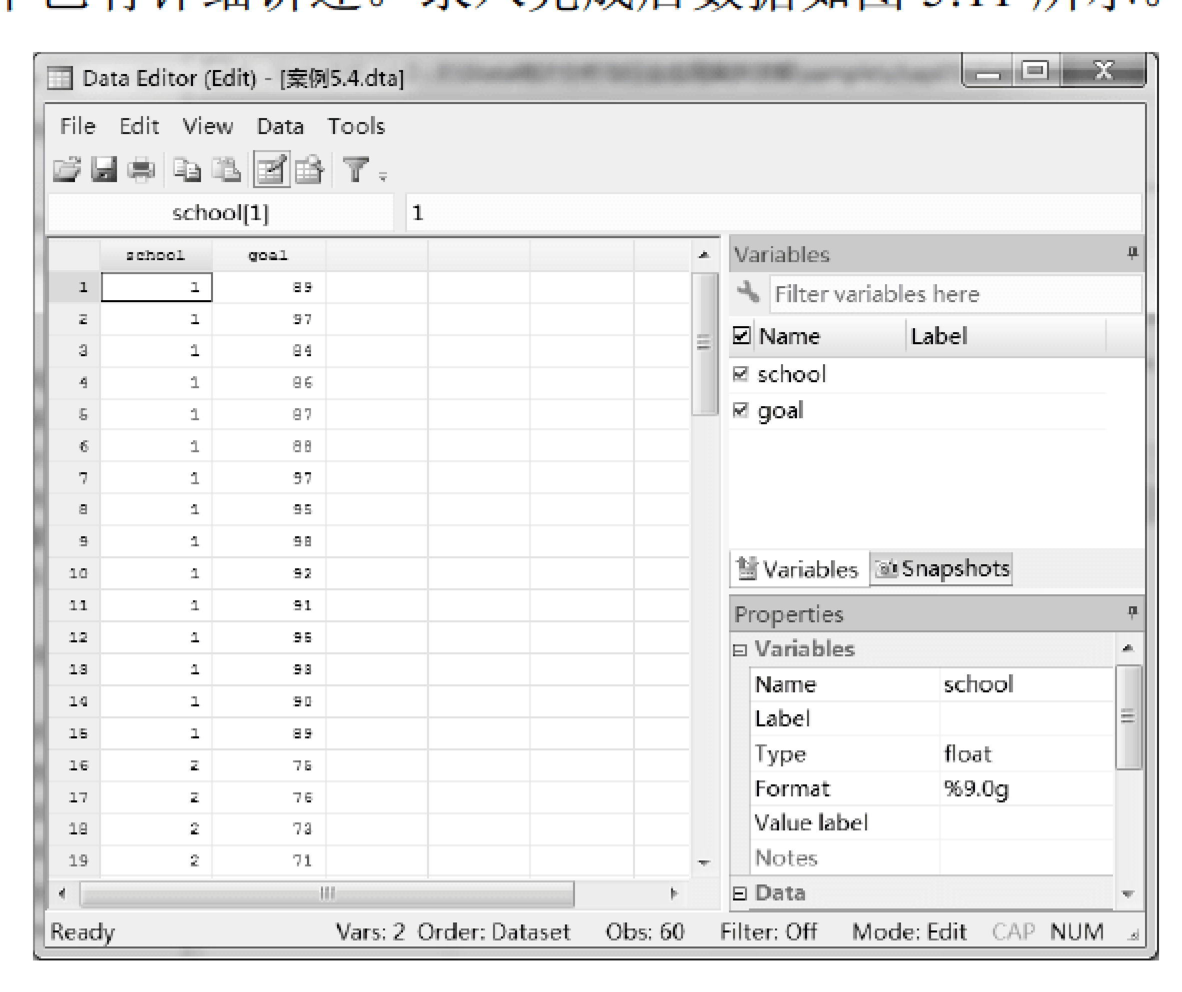


图 5.11　案例 5.4 数据

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入如下命令（旨在用多独立样本检验方法分析毕业于不同高校的员工在工作上的表现是否有显著的差异）：

```
kwallis  goal,by( school)
```

03 设置完毕后，按键盘上的回车键，等待输出结果。

5.4.4 结果分析

在 Stata 14.0 主界面的结果窗口我们可以看到如图 5.12 所示的分析结果。

通过观察分析结果，我们可以看出有 4 组，每组有 15 个，共有 60 个有效样本参与了假设检验，p 值远小于 0.05，所以需要拒绝原假设，也就是说，毕业于不同高校的员工在工作上的表现有显著的差异。

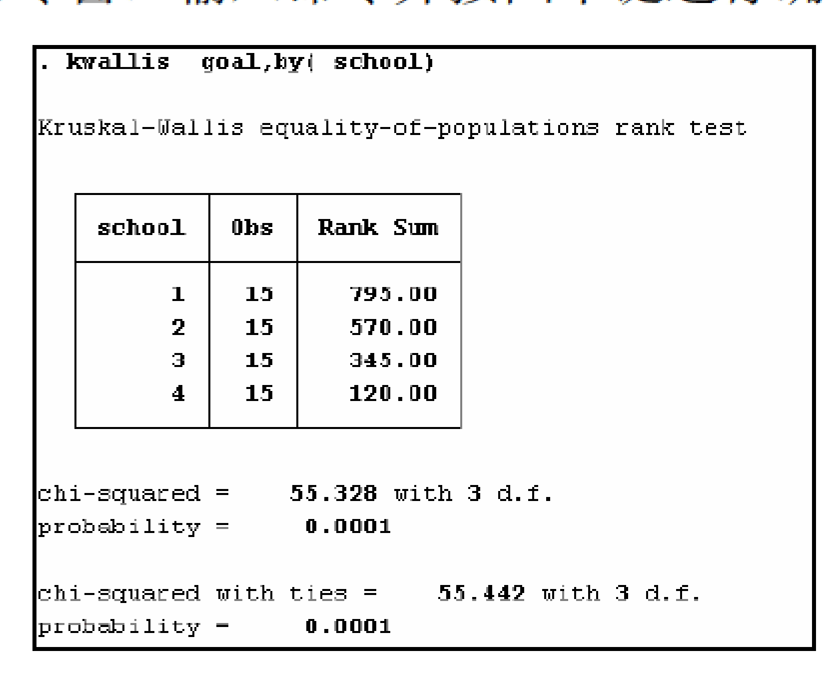
5.4.5 案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

例如，我们只针对 goal 变量大于 75 的观测样本进行多独立样本检验，那么操作命令即为：

```
kwallis  goal if goal>75,by( school)
```

在命令窗口输入命令并按回车键进行确认，结果如图 5.13 所示。



```
. kwallis  goal,by( school)

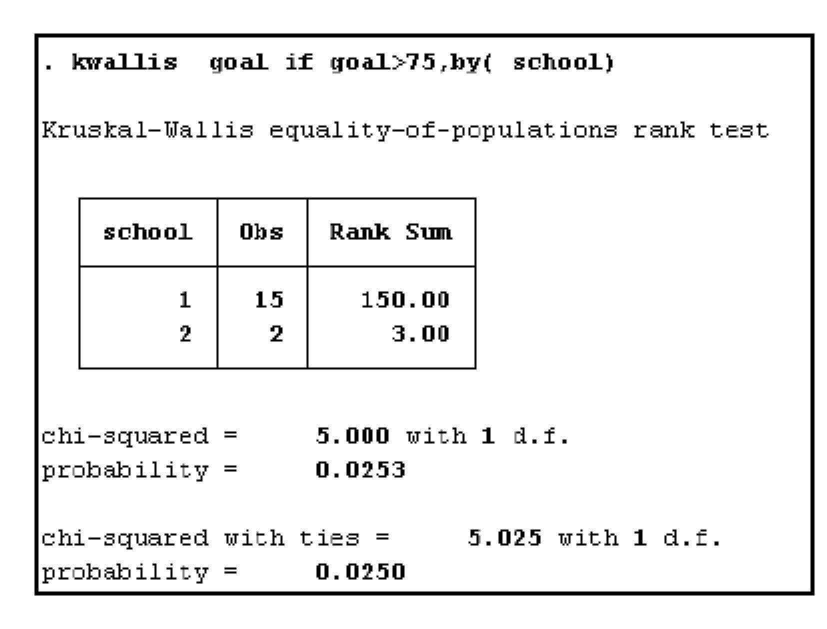
Kruskal-Wallis equality-of-populations rank test

  school   Obs   Rank Sum
       1    15     795.00
       2    15     570.00
       3    15     345.00
       4    15     120.00

chi-squared =     55.328 with 3 d.f.
probability =      0.0001

chi-squared with ties =     55.442 with 3 d.f.
probability =      0.0001
```

图 5.12　分析结果图



```
. kwallis  goal if goal>75,by( school)

Kruskal-Wallis equality-of-populations rank test

  school   Obs   Rank Sum
       1    15     150.00
       2     2       3.00

chi-squared =      5.000 with 1 d.f.
probability =      0.0253

chi-squared with ties =      5.025 with 1 d.f.
probability =      0.0250
```

图 5.13　分析结果图

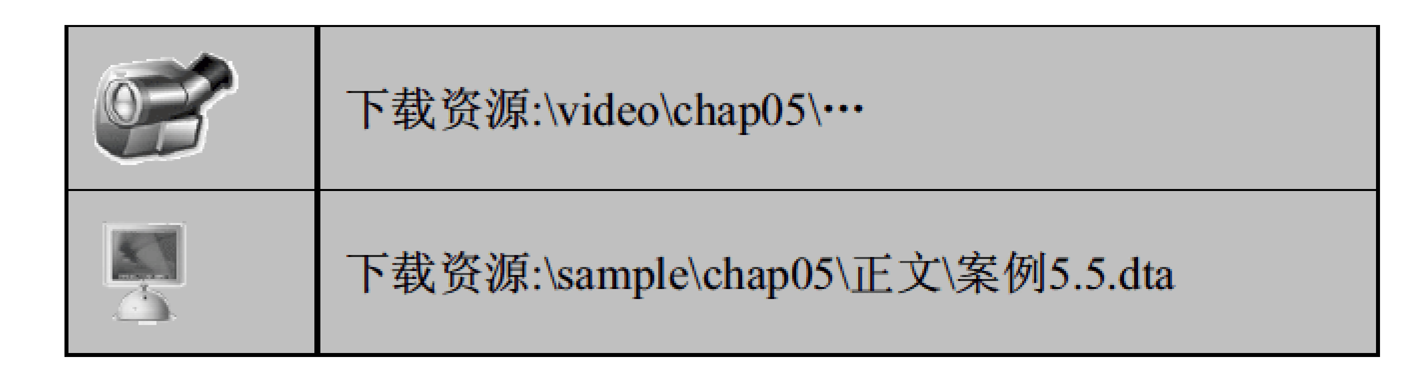
通过观察分析结果，我们可以看出参与分析的样本由 4 组变为 2 组，共有 17 个有效样本参与了假设检验，p 值远小于 0.05，所以需要拒绝原假设。

5.5　实例五——游程检验

5.5.1 游程检验的功能与意义

Stata 的游程检验（Runs Test）也是非参数检验方法的一种，其基本功能是：可以判断样本序列是否为随机序列。这种检验过程是通过分析游程的总个数来实现的。

5.5.2　相关数据来源



【例 5.5】表 5.5 给出了某纺织厂连续 15 天通过试验得出的 28 号梳棉棉条的棉结杂质粒数的数据。试用游程检验方法研究该纺织厂的生产情况是否正常。

表 5.5　棉结杂质粒数表

天数编号	棉结杂质粒数/粒/g
001	52
002	89
003	45
004	75
005	62
006	64
007	64
008	62
009	65
010	65
011	64
012	38
013	51
014	46
015	78

5.5.3　Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中只有一个变量，即棉结杂质粒数。我们把棉结杂质粒数变量设定为 number，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作我们在第 1 章中已有详细讲述。录入完成后数据如图 5.14 所示。

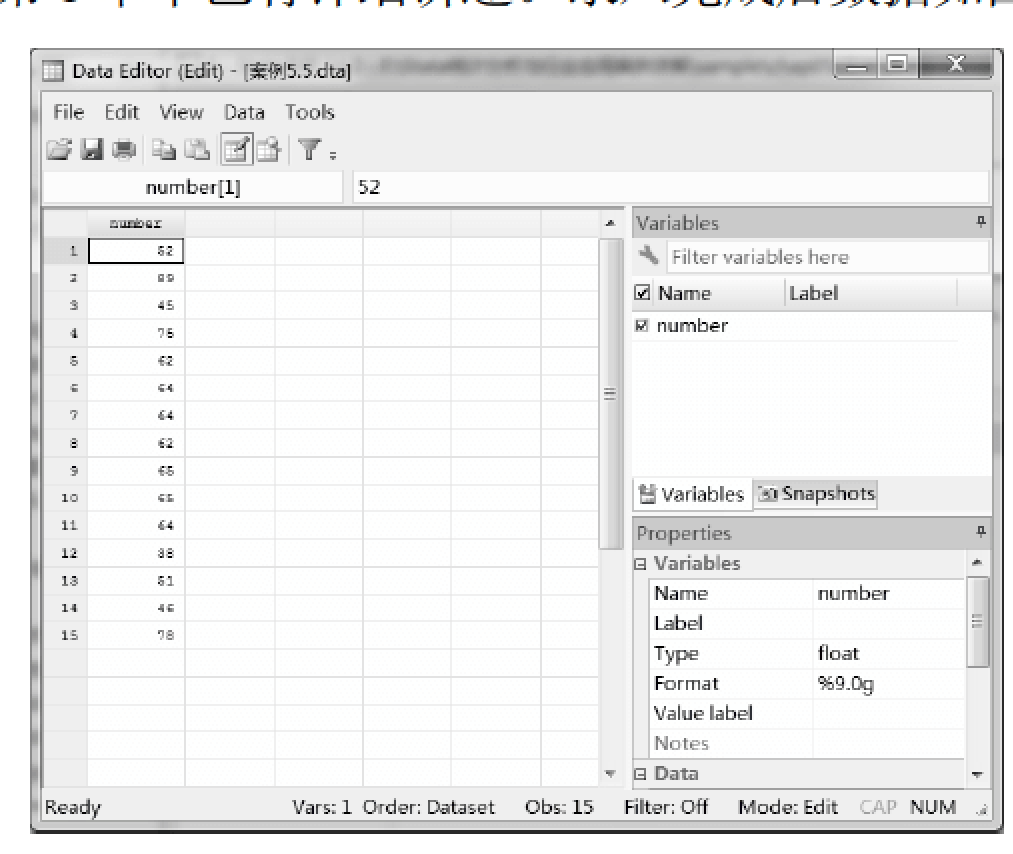


图 5.14　案例 5.5 数据

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入如下命令（本命令的含义是判断 number 变量是否为随机）：

```
runtest  number
```

03 设置完毕后，按键盘上的回车键，等待输出结果。

5.5.4 结果分析

在 Stata 14.0 主界面的结果窗口我们可以看到如图 5.15 所示的分析结果。

```
. runtest  number
 N(number <= 64) = 10
 N(number >  64) = 5
            obs = 15
        N(runs) = 8
             z  = .2
       Prob>|z| = .84
```

图 5.15　分析结果图

通过观察分析结果，我们可以看出 Prob>|z| = 0.84，远大于 0.05，所以需要接受原假设，也就是说，数据的产生是随机的，不存在自相关现象，该纺织厂的生产情况正常。

5.5.5 案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

Stata 14.0 默认采用中位数作为参考值，如果设定均值作为参考值，那么操作命令即为：

```
runtest  number,mean
```

在命令窗口输入命令并按回车键进行确认，结果如图 5.16 所示。

```
. runtest  number,mean
 N(number <= 61.33333333333334) = 5
 N(number >  61.33333333333334) = 10
            obs = 15
        N(runs) = 6
             z  = -1.01
       Prob>|z| = .31
```

图 5.16　分析结果图

通过观察分析结果，我们可以看出 Prob>|z| = 0.31，远大于 0.05，所以需要接受原假设，也就是说，数据的产生是随机的，不存在自相关现象。

5.6　本章习题

（1）表 5.6 给出了某实验中学 60 名毕业生的高考数学成绩。试用单样本正态分布检验方法研究其是否服从正态分布。

表 5.6　某实验中学 60 名毕业生的高考数学成绩

编号	高考数学成绩
001	144
002	142
003	141
004	138
005	129
…	…
058	126
059	128
060	134

（2）表 5.7 给出了 A、B 两家公司近些年的净利润情况。试用两独立样本检验方法判断两家公司近些年的净利润是否存在显著差异。

（3）为了研究一种智力开发课程的效果，特抽取了 30 名学生进行试验，其使用该产品前后的智商如表 5.8 所示。试用配对样本 T 检验的方法判断该开发课程是否有效。

表 5.7　A、B 两家公司近些年的净利润（单位：万元）

年份	净利润	
	A公司	B公司
2001	1461.7	1586.1
2002	1607.8	1726.9
2003	1709.0	1284.8
2004	1171.0	1766.4
2005	1361.5	1693.1
2006	1847.5	1815.3
2007	1458.2	1737.5
2008	1033.8	1318.7
2009	1850.9	1318.2
2010	1560.3	1889.2
2011	1110.3	1480.9
2012	1415.2	1251.8

表 5.8　使用智力开发课程前后的智商水平

编号	使用智力开发课程前	使用智力开发课程后
001	121	123
002	86	88
003	97	99
004	102	103
005	104	105

（续表）

编号	使用智力开发课程前	使用智力开发课程后
…	…	…
028	93	101
029	86	95
030	87	99

（4）参加某足球俱乐部试训的一批球员来自 4 个不同的国家，从而来源于 4 个不同国家的球员构成了 4 个独立的样本。试训期结束后，教练员对这些球员进行考察打分，结果如表 5.9 所示。试用多独立样本检验方法分析来自于不同国家的球员表现是否有显著的差异。

表 5.9　球员考核成绩

A国	87	79	94	91	89	85	77
B国	67	69	72	75	76	69	79
C国	58	48	50	49	36	50	42
D国	20	29	39	38	29	20	15

（5）表 5.10 给出了某汽车连续 15 天每加仑汽油行驶的英里数。试用游程检验方法研究该汽车每加仑汽油行驶英里数是否为随机。

表 5.10　每加仑汽油行驶英里数

天数编号	每加仑汽油行驶英里数
001	18.4
002	17.5
003	16.0
004	16.9
005	20.5
006	22.4
007	21.4
008	20.6
009	19.5
010	23.1
011	21.3
012	22.9
013	22.5
014	20.1
015	19.1

第 6 章　Stata 方差分析

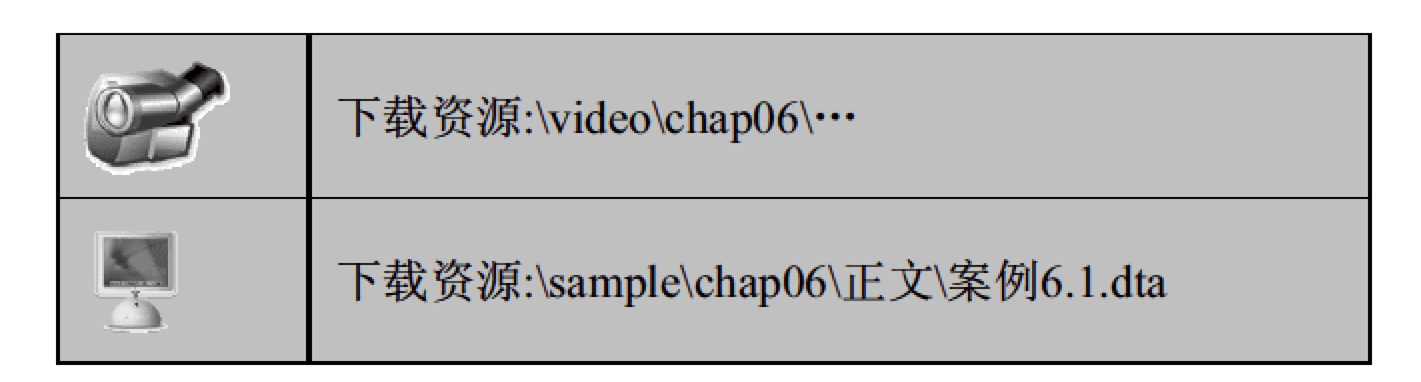
当遇到多个平均数间的差异显著性检验时，我们可以采用方差分析法。方差分析法就是将所要处理的观测值作为一个整体，按照变异的不同来源把观测值总变异的平方和以及自由度分解为两个或多个部分，从而获得不同变异来源的均方与误差均方；通过比较不同变异来源的均方与误差均方，判断各样本所属总体方差是否相等。方差分析主要包括单因素方差分析、多因素方差分析、协方差分析、重复测量方差分析等。下面我们将分别介绍这些方法在实例中的应用。

6.1　实例一——单因素方差分析

6.1.1　单因素方差分析的功能与意义

单因素方差分析是方差分析（Analysis of Variance）类型中最基本的一种，研究的是一个因素对于试验结果的影响和作用，这一因素可以有不同的取值或者是分组。单因素方差分析所要检验的问题就是当因素选择不同的取值或者分组时对结果有无显著的影响。

6.1.2　相关数据来源



【例 6.1】表 6.1 给出了 4 种新型药物对白鼠胰岛素分泌水平的影响测量结果，数据为白鼠的胰岛质量。试用单因素方差分析检验 4 种药物对胰岛素水平的影响是否相同。

表 6.1　4 种药物刺激下的白鼠胰岛质量

测量编号	胰岛质量/g	药物组
1	86.1	1
2	89.5	1
3	71.5	1
4	86.2	1
5	85.7	1
6	82.7	1
…	…	…
36	86.4	4

（续表）

测量编号	胰岛质量/g	药物组
37	86.4	4
38	87	4
39	86	4
40	88.3	4

6.1.3 Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有两个变量，分别为胰岛质量和药物组。我们把胰岛质量变量设定为 weight，把药物组变量设定为 group，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作我们在第 1 章中已有详细讲述。录入完成后数据如图 6.1 所示。

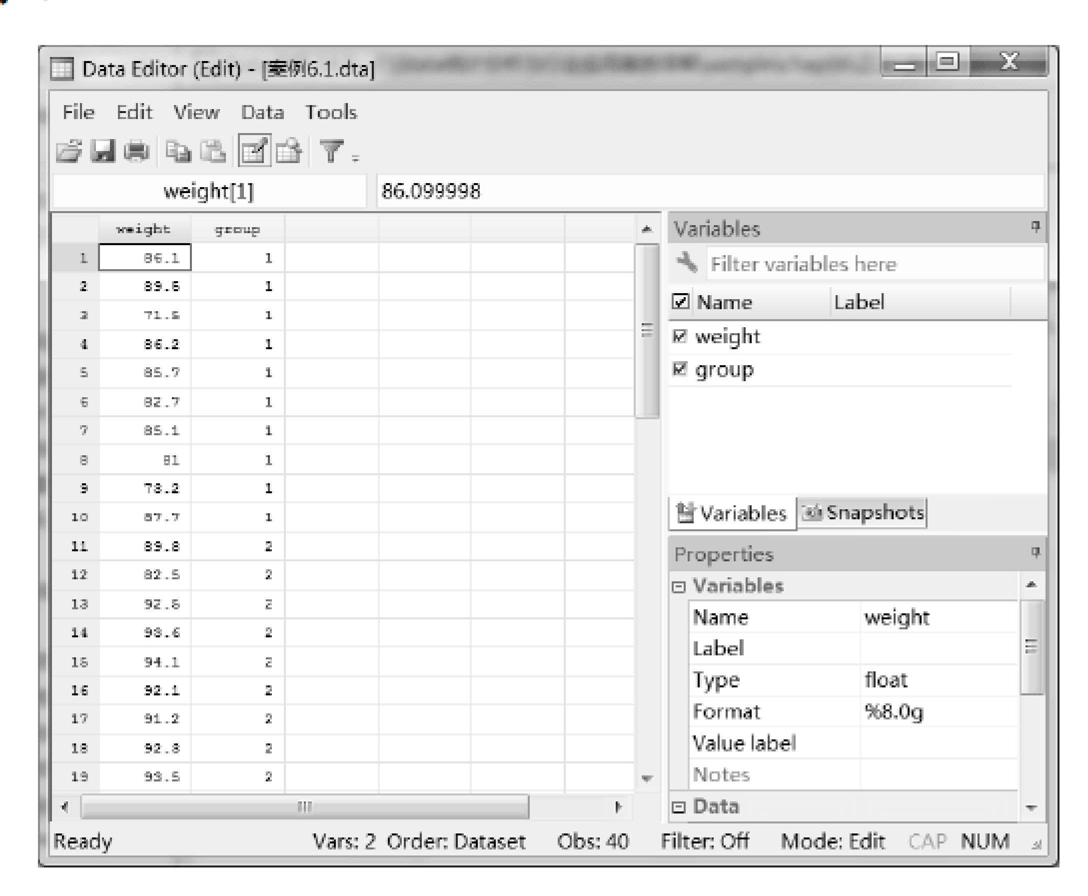


图 6.1　案例 6.1 数据

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入如下命令（旨在用单因素方差分析检验 4 种药物对胰岛素水平的影响是否相同）：

```
oneway weight group, tabulate
```

03 设置完毕后，按键盘上的回车键，等待输出结果。

6.1.4 结果分析

在 Stata 14.0 主界面的结果窗口我们可以看到如图 6.2 所示的分析结果。

```
. oneway  weight group,tabulate

            |       Summary of weight
      group |        Mean   Std. Dev.       Freq.
------------+------------------------------------
          1 |   82.869998   6.0378526          10
          2 |       91.58   3.4701259          10
          3 |       73.42   1.5389754          10
          4 |   85.830001   1.7550251          10
------------+------------------------------------
      Total |      83.425   7.5319406          40

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      1726.96106      3   575.653686     42.68     0.0000
 Within groups      485.513964     36    13.486499
------------------------------------------------------------------------
    Total           2212.47502     39   56.7301288

Bartlett's test for equal variances:  chi2(3) =  20.0858  Prob>chi2 = 0.000
```

图 6.2　分析结果图

从上述分析结果中可以得到很多信息。分析结果图的上半部分是胰岛质量变量的概要统计，其中共有 4 个组别，第 1 组的均值是 82.869998，标准差是 6.0378526，频数是 10；第 2 组的均值是 91.58，标准差是 3.4701259，频数是 10；第 3 组的均值是 73.42，标准差是 1.5389754，频数是 10；第 4 组的均值是 85.830001，标准差是 1.7550251，频数是 10。样本总数是 40 个，均值是 83.425，标准差是 7.5319406。下半部分是方差分析的结果，chi2(3) = 20.0858，Prob>chi2 = 0.000，说明要拒绝等方差假设，也就是说本例的结论是 4 种药物对胰岛素水平的影响显著不相同。

6.1.5　案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

例如，我们只针对 weight 变量大于 72 的观测样本进行单因素方差分析，那么操作命令即为：

```
oneway weight group if weight>72, tabulate
```

在命令窗口输入命令并按回车键进行确认，结果如图 6.3 所示。

```
. oneway weight group if weight>72, tabulate

            |       Summary of weight
      group |        Mean   Std. Dev.       Freq.
------------+------------------------------------
          1 |   84.133331   4.8018229           9
          2 |       91.58   3.4701259          10
          3 |   73.862499   1.3752285           8
          4 |   85.830001   1.7550251          10
------------+------------------------------------
      Total |   84.383783   6.9894969          37

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      1424.91462      3   474.971541     46.96     0.0000
 Within groups      333.795779     33   10.1150236
------------------------------------------------------------------------
    Total            1758.7104     36   48.8530667

Bartlett's test for equal variances:  chi2(3) =  13.5840  Prob>chi2 = 0.004
```

图 6.3　分析结果图

对该结果的详细说明在前面已有提及，此处限于篇幅不再赘述。chi2(3) = 13.5840，Prob>chi2 = 0.004，说明要拒绝等方差假设。

6.2　实例二——多因素方差分析

6.2.1　多因素方差分析的功能与意义

多因素方差分析的基本思想基本等同于单因素方差分析，不同之处在于其研究的是两个或者两个以上因素对于试验结果的作用和影响，以及这些因素共同作用的影响。多因素方差分析所要研究的是多个因素的变化是否会导致试验结果的变化。由于三因素以及三因素以上方差分析较少用到，因此下面我们以双因素方差分析为例进行介绍。

6.2.2　相关数据来源

	下载资源:\video\chap06\…
	下载资源:\sample\chap06\正文\案例6.2.dta

【例 6.2】将 40 只大鼠随机等分为 4 组，每组 10 只，进行肌肉损伤后的缝合试验。处理方式由两个因素组合而成，A 因素为缝合方法，分别为外膜缝合和内膜缝合，记作 a1、a2；B 因素为缝合后的时间，分别为缝合后 1 月和 2 月，记作 b1、b2。试验结果为大鼠肌肉缝合后肌肉力度的恢复度（%），如表 6.2 所示，从而考察缝合方法和缝合后时间对肌肉力度的恢复度是否有显著影响。

表 6.2　大鼠肌肉缝合后肌肉力度的恢复度测量数据

测量编号	肌肉力度的恢复度/%	缝合方法	缝合后时间
1	10.5	a1	b1
2	10.6	a1	b1
3	11.5	a1	b1
4	11.3	a1	b1
5	11	a1	b1
6	11.4	a1	b1
…	…	…	…
38	28.3	a2	b2
39	28.1	a2	b2
40	28.3	a2	b2

6.2.3　Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有 3 个变量，分别是肌

肉力度的恢复度、缝合方法和缝合后时间。我们把肌肉力度的恢复度变量设定为 renew，把缝合方法变量设定为 method，并且其中的缝合方法 a1 设定为 1、缝合方法 a2 设定为 2，把缝合后时间变量设定为 time，并且其中的缝合方法 b1 设定为 1、缝合方法 b2 设定为 2，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作我们在第 1 章中已有详细讲述。录入完成后数据如图 6.4 所示。

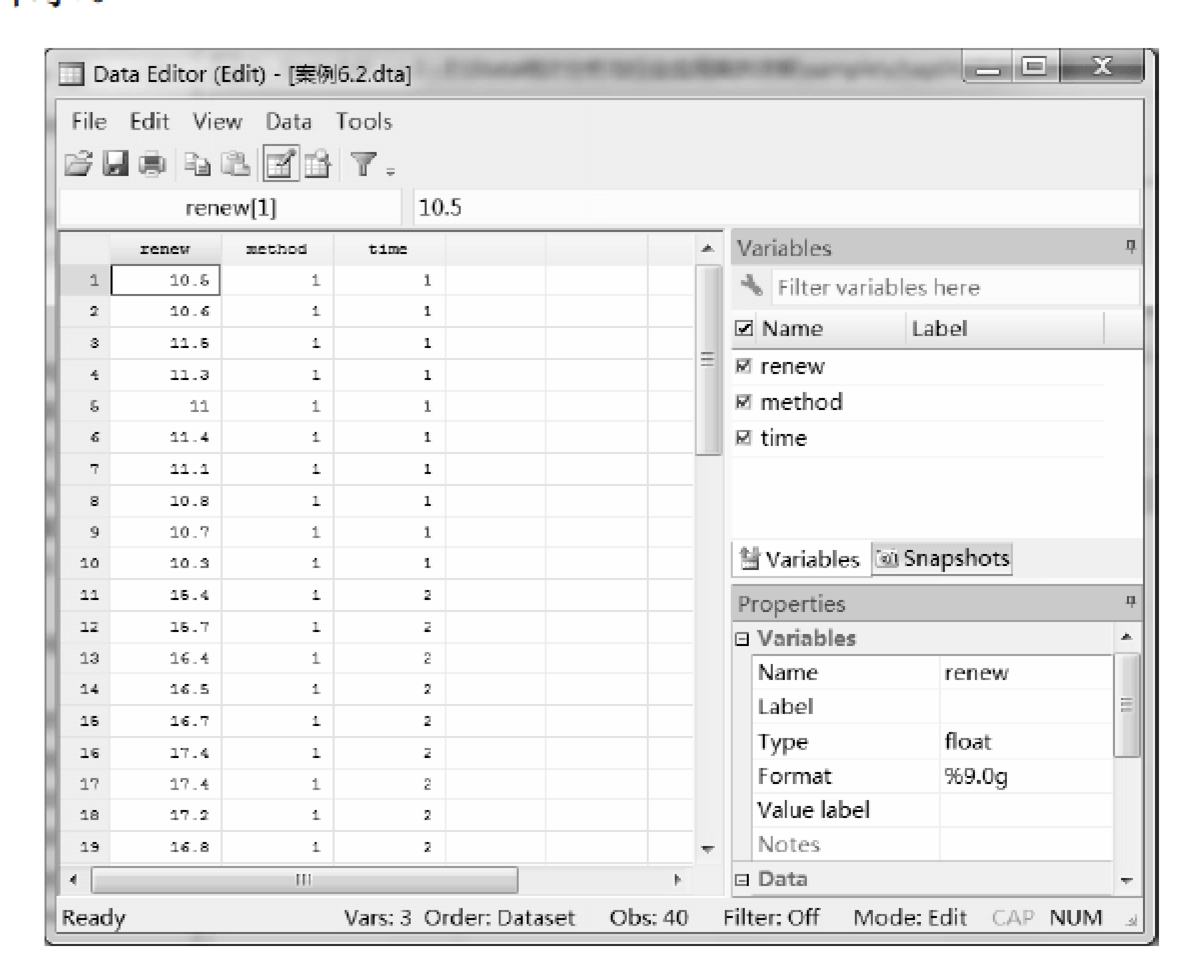


图 6.4　案例 6.2 数据

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入如下命令（旨在考察缝合方法和缝合后时间对肌肉力度的恢复度是否有显著影响）：

```
anova  renew method time method# time
```

03 设置完毕后，按键盘上的回车键，等待输出结果。

6.2.4　结果分析

在 Stata 14.0 主界面的结果窗口我们可以看到如图 6.5 所示的分析结果。

```
. anova  renew method time method# time

                         Number of obs =         40    R-squared     =  0.9941
                         Root MSE      =    .516774    Adj R-squared =  0.9936

                  Source | Partial SS    df       MS           F     Prob > F
              -----------+----------------------------------------------------
                   Model |  1617.92495     3  539.308318    2019.46     0.0000
                         |
                  method |  1322.49997     1  1322.49997    4952.15     0.0000
                    time |  294.848987     1  294.848987    1104.07     0.0000
             method#time |  .575999588     1  .575999588       2.16     0.1506
                         |
                Residual |  9.61400039    36  .267055566
              -----------+----------------------------------------------------
                   Total |  1627.53895    39   41.731768
```

图 6.5　分析结果图

通过观察分析结果，我们可以看出共有 40 个有效样本参与了方差分析。

- 可决系数（R-squared）以及修正的可决系数（Adj R-squared）都非常接近于 1，这说明模型的拟合程度很高，也就是说模型的解释能力很强。
- Prob > F Model=0.0000，说明模型的整体是很显著的。
- Prob > F method =0.0000，说明变量 method 的主效应是非常显著的。
- Prob > F time =0.0000，说明变量 time 的主效应也是非常显著的。
- Prob > F method#time = 0.1506，说明变量 method 与变量 time 的交互效应是不显著的。这一点也可以从下面的命令中得到验证。

在主界面的“Command”文本框中分别输入下列命令并按键盘上的回车键：

```
test method
test time
test method#time
```

可以得到如图 6.6 所示的结果。

```
test  method

                  Source   Partial SS    df       MS            F     Prob > F

                  method   1322.49997     1   1322.49997     4952.15    0.0000
                Residual   9.61400039    36   .267055566

test  time

                  Source   Partial SS    df       MS            F     Prob > F

                    time   294.848987     1   294.848987     1104.07    0.0000
                Residual   9.61400039    36   .267055566

test  method# time

                  Source   Partial SS    df       MS            F     Prob > F

             method#time   .575999588     1   .575999588        2.16    0.1506
                Residual   9.61400039    36   .267055566
```

图 6.6　分析结果图

在上面的例子中，因为变量 method 与变量 time 的交互效应是不显著的，所以我们可以构建更加简单的不包含两者交互效应的方差分析模型。在主界面的“Command”文本框中输入下列命令并按键盘上的回车键：

```
anova  renew method time
```

可以得到如图 6.7 所示的结果。

```
. anova  renew method time

                        Number of obs =         40    R-squared     =  0.9937
                        Root MSE      =    .524791    Adj R-squared =  0.9934

                  Source   Partial SS    df       MS            F     Prob > F

                   Model   1617.34895     2   808.674477     2936.31    0.0000

                  method   1322.49997     1   1322.49997     4802.01    0.0000
                    time   294.848987     1   294.848987     1070.60    0.0000

                Residual        10.19    37   .275405405

                   Total   1627.53895    39    41.731768
```

图 6.7　分析结果图

至此，我们以两个因素介绍了多因素方差分析的应用。事实上，多因素方差分析的模型构建是非常灵活的，如果存在 3 个或者 3 个因素以上，我们要纳入任何一项变量间的交互效应，则只需指定有关变量名称，并且之间用“#”连接（注意，之前的很多 Stata 版本用的是“*”）即可。

6.2.5 案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

例如，我们只针对 renew 变量大于 11 的观测样本进行多因素方差分析，那么操作命令即为：

```
anova  renew method time method# time if renew>11
```

在命令窗口输入命令并按回车键进行确认，结果如图 6.8 所示。

```
. anova  renew method time method# time if renew>11

                           Number of obs =        34    R-squared     =  0.9923
                           Root MSE      =   .523625    Adj R-squared =  0.9916

                  Source | Partial SS    df       MS           F    Prob > F
             ------------+----------------------------------------------------
                   Model | 1065.52889     3   355.176296    1295.40    0.0000
                         |
                  method | 928.243661     1   928.243661    3385.49    0.0000
                    time | 198.740037     1   198.740037     724.84    0.0000
             method#time | .010227234     1   .010227234       0.04    0.8482
                         |
                Residual | 8.22550068    30   .274183356
             ------------+----------------------------------------------------
                   Total | 1073.75439    33   32.5380118
```

图 6.8 分析结果图

通过观察分析结果，我们可以看出共有 34 个有效样本参与了方差分析。

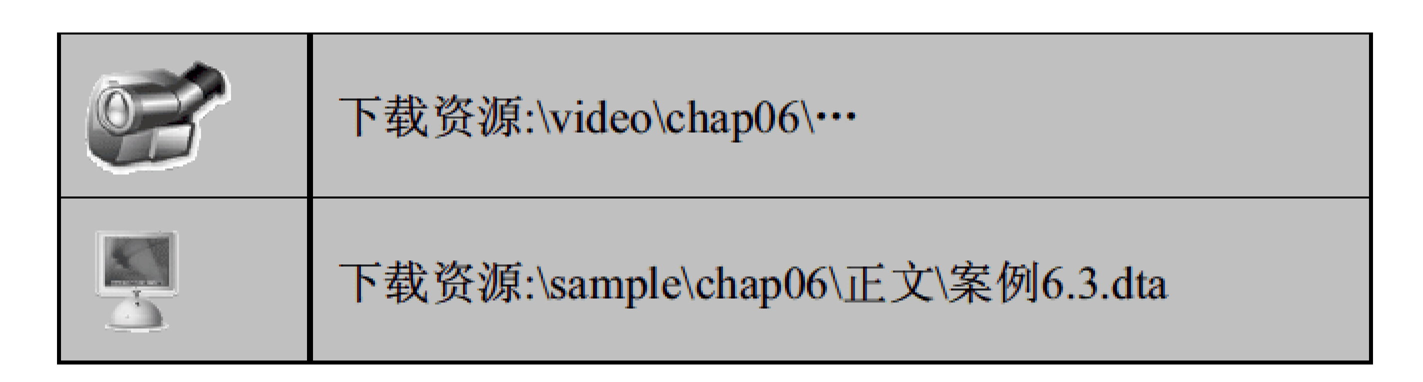
- 可决系数（R-squared）以及修正的可决系数（Adj R-squared）都非常接近于 1，这说明模型的拟合程度很高，也就是说模型的解释能力很强。
- Prob > F Model=0.0000，说明模型的整体是很显著的。
- Prob > F method =0.0000，说明变量 method 的主效应是非常显著的。
- Prob > F time =0.0000，说明变量 time 的主效应也是非常显著的。
- Prob > F method#time = 0.8482，说明变量 method 与变量 time 的交互效应是不显著的。

6.3 实例三——协方差分析

6.3.1 协方差分析的功能与意义

协方差分析是将回归分析同方差分析结合起来，以消除混杂因素的影响，是对试验数据进行分析的一种分析方法。一般情况下，协方差分析研究比较一个或者几个因素在不同水平上的差异，但观测量同时还受另一个难以控制的协变量的影响，在分析中剔除其影响，再分析各因素对观测变量的影响。

6.3.2 相关数据来源



【例 6.3】某学校实施新政策以改善部分年轻教师的生活水平。政策实施后开始对年轻教师待遇的改善情况进行调查，调查结果如表 6.3 所示。用实施新政策后的工资来反映生活水平的提高，要求剔除实施新政策前的工资差异，试分析教师的级别和该新政策对年轻教师工资的提高是否有显著的影响。

表 6.3　年轻教师工资表（单位：千元）

年龄	原工资	现工资	教师级别	政策实施
26	4	5	2	否
27	3	4	3	否
27	3	5	1	是
29	2	4	2	否
28	5	6	2	是
…	…	…	…	…
29	6	9	3	是
27	8	10	2	否

6.3.3 Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有 5 个变量，分别为年龄、原工资、现工资、教师级别和政策实施。我们把年龄这一变量设定为 age，把原工资这一变量设定为 beforesalary，把现工资这一变量设定为 nowsalary，把教师级别这一变量设定为 identity，把政策实施这一变量设定为 policy，并且用“1”表示“实施政策”，而用“0”表示“没有实施政策”，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作我们在第 1 章中已有详细讲述。录入完成后数据如图 6.9 所示。

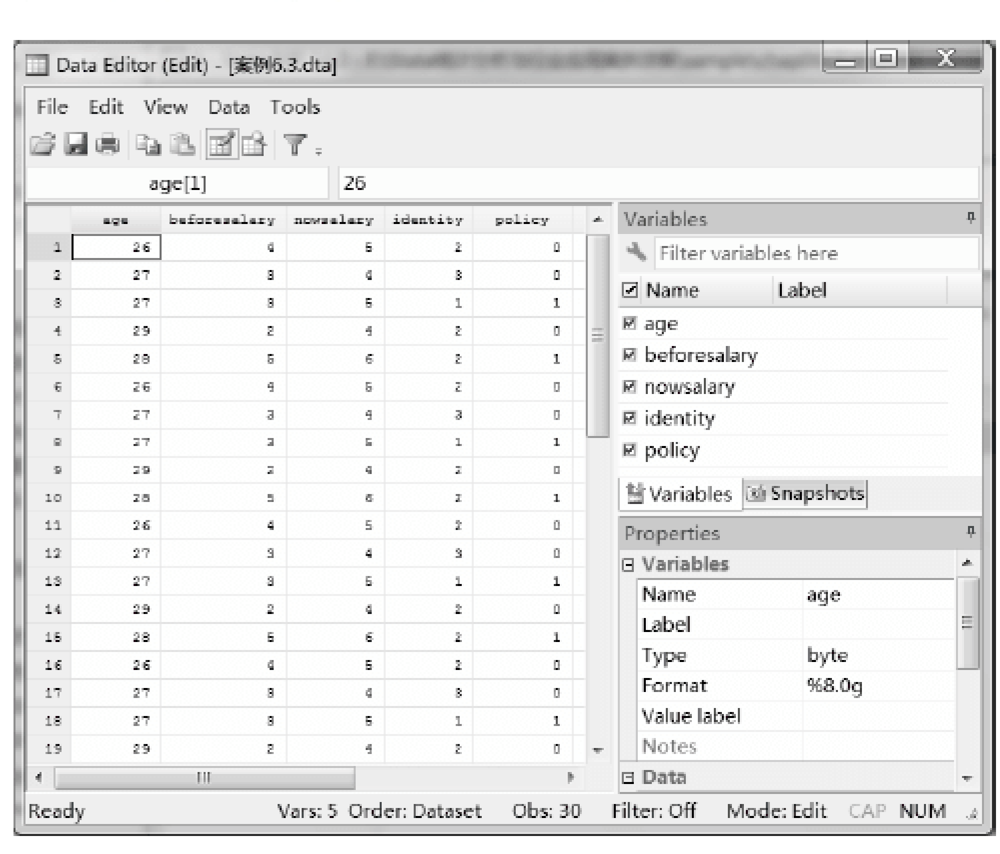


图 6.9　案例 6.3 数据

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入如下命令（旨在分析教师的级别和新政策对年轻教师工资的提高是否有显著的影响）：

```
anova  nowsalary identity policy c.beforesalary
```



c.beforesalary 的意义是说明 beforesalary 是一个连续变量，在一些 Stata 旧版本中，本例的命令应该是：anova　nowsalary identity policy,continuous(beforesalary)。

03 设置完毕后，按键盘上的回车键，等待输出结果。

6.3.4　结果分析

在 Stata 14.0 主界面的结果窗口我们可以看到如图 6.10 所示的分析结果。

```
. anova  nowsalary identity policy c.beforesalary

                         Number of obs =         30    R-squared     =  0.8705
                         Root MSE      =    .547489    Adj R-squared =  0.8498

                  Source | Partial SS    df       MS           F    Prob > F
              -----------+----------------------------------------------------
                   Model |  50.3730714     4  12.5932679     42.01    0.0000
                         |
                identity |  .905719977     2  .452859989      1.51    0.2402
                  policy |  .002217987     1  .002217987      0.01    0.9321
            beforesal~y  |  34.0025734     1  34.0025734    113.44    0.0000
                         |
                Residual |  7.49359522    25  .299743809
              -----------+----------------------------------------------------
                   Total |  57.8666667    29   1.9954023
```

图 6.10　分析结果图

通过观察分析结果，我们可以看出共有 30 个有效样本参与了方差分析。

- 可决系数（R-squared）以及修正的可决系数（Adj R-squared）都超过了 80%，这说明模型的拟合程度很高，也就是说模型的解释能力很强。
- Prob＞F Model=0.0000，说明模型的整体是很显著的。
- Prob＞F identity =0.2402，说明变量 identity 的主效应是非常不显著的。
- Prob＞F policy =0.9321，说明变量 policy 的主效应也是非常不显著的。
- Prob＞F beforesalary =0.0000，说明变量 beforesalary 的主效应是非常显著的。

也就是说，教师的级别和新政策是否实施对年轻教师工资的提高都没有显著的影响，而实施新政策前的工资差异是对年轻教师的现有工资有显著影响的。

在此基础上，我们可以对模型进行改进，即引入变量的交互项进行深入分析，我们在主界面的“Command”文本框中分别输入下列命令并按键盘上的回车键：

```
anova  nowsalary identity policy c.beforesalary c.beforesalary# identity
c.beforesalary# policy identity# policy
```

可以得到如图 6.11 所示的结果。

```
. anova  nowsalary identity policy c.beforesalary  c.beforesalary# identity   c.beforesalary# policy i
> dentity# policy

                              Number of obs =         30    R-squared      =  0.9551
                              Root MSE      =    .328897    Adj R-squared =  0.9458

                    Source | Partial SS    df       MS           F     Prob > F
              -------------+----------------------------------------------------
                     Model | 55.2705128     5  11.0541026     102.19    0.0000
                           |
                  identity | 5.36344525     2  2.68172262      24.79    0.0000
                    policy | .492470492     1  .492470492       4.55    0.0433
               beforesal~y |  31.840922     1   31.840922     294.35    0.0000
      identity#beforesal~y | 4.89744137     1  4.89744137      45.27    0.0000
        policy#beforesal~y |          0     0
           identity#policy |          0     0
                           |
                  Residual | 2.59615385    24  .108173077
              -------------+----------------------------------------------------
                     Total | 57.8666667    29   1.9954023
```

图 6.11　分析结果图

在本分析结果中，我们可以看到 c.beforesalary# policy identity# policy 这两个交互项是不起作用的，所以我们要把它们去掉，在主界面的“Command”文本框中分别输入下列命令并按键盘上的回车键：

```
anova  nowsalary identity policy c.beforesalary c.beforesalary# identity
```

可以得到如图 6.12 所示的结果。

```
. anova  nowsalary identity policy c.beforesalary  c.beforesalary# identity

                              Number of obs =         30    R-squared      =  0.9551
                              Root MSE      =    .328897    Adj R-squared =  0.9458

                    Source | Partial SS    df       MS           F     Prob > F
              -------------+----------------------------------------------------
                     Model | 55.2705128     5  11.0541026     102.19    0.0000
                           |
                  identity | 5.36344525     2  2.68172262      24.79    0.0000
                    policy | .492470492     1  .492470492       4.55    0.0433
               beforesal~y |  31.840922     1   31.840922     294.35    0.0000
      identity#beforesal~y | 4.89744137     1  4.89744137      45.27    0.0000
                           |
                  Residual | 2.59615385    24  .108173077
              -------------+----------------------------------------------------
                     Total | 57.8666667    29   1.9954023
```

图 6.12　分析结果图

通过观察本分析结果，我们可以看出：

- 可决系数（R-squared）以及修正的可决系数（Adj R-squared）得到进一步提高，超过了 90%，说明模型的拟合程度得到了进一步提高，也就是说模型的解释能力变强了。
- Prob > F Model=0.0000，说明模型的整体是很显著的。
- Prob > F identity =0.0000，说明变量 identity 的主效应是非常显著的。
- Prob > F policy =0.0433，说明变量 policy 的主效应也是显著的。
- Prob > F beforesalary =0.0000，说明变量 beforesalary 的主效应是非常显著的。
- Prob > F c.beforesalary# identity =0.0000，说明变量 beforesalary 与 identity 的交互效应是非常显著的。

也就是说，教师的级别、新政策是否实施、实施新政策前的工资差异都对年轻教师的现

有工资有显著影响，教师的级别与实施新政策前的工资差异的交互效应也对年轻教师的现有工资有显著影响。

此外，我们可以针对这一结果进行回归分析，在主界面的“Command”文本框中输入下列命令并按键盘上的回车键：

```
regress
```

可以得到如图 6.13 所示的结果。

```
. regress

      Source |       SS           df       MS      Number of obs   =        30
-------------+----------------------------------   F(  5,     24)  =    102.19
       Model |  55.2705128         5  11.0541026   Prob > F        =    0.0000
    Residual |  2.59615385        24  .108173077   R-squared       =    0.9551
-------------+----------------------------------   Adj R-squared   =    0.9458
       Total |  57.8666667        29   1.9954023   Root MSE        =     .3289

------------------------------------------------------------------------------
   nowsalary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    identity |
          2  |   1.903846   .4334928     4.39   0.000     1.009161    2.798531
          3  |  -1.423077   .2745441    -5.18   0.000    -1.989708   -.8564458
             |
    1.policy |  -.4230769   .1982845    -2.13   0.043    -.8323161   -.0138378
beforesalary |   1.807692   .1356133    13.33   0.000       1.5278    2.087584
             |
identity#c.beforesalary |
          2  |  -.9038462   .1343289    -6.73   0.000    -1.181087   -.6266049
          3  |          0  (omitted)
             |
       _cons |  -5.95e-14   .3797773    -0.00   1.000    -.7838217    .7838217
------------------------------------------------------------------------------
```

图 6.13　分析结果图

在这个结果中，我们可以发现前面的实例相当于把 nowsalary 这一变量作为因变量，把 identity、policy、beforesalary、beforesalary 与 identity 的交互项这 4 个变量作为自变量进行了一次回归分析。系统针对每个分类自变量（包括 identity、policy 以及 beforesalary 与 identity 的交互项）创建了相应的虚拟变量，这里要把单个虚拟变量的回归系数理解为它对因变量的预测值或者条件平均数的效应。例如，1.policy 表示那些具有同样教师级别以及同样改革前工资的年轻教师中，接受新政策改革的现有工资要比没有接受新政策改革的低 42.30769 个百分点。此外，我们还得到了每个系数的置信区间和单项 T 检验的结果，相比于单纯的方差分析，我们从这一结果中得到的信息要丰富得多。

6.3.5　案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

例如，我们只针对 age 变量大于 26 的观测样本进行协方差分析，那么操作命令即为：

```
anova  nowsalary identity policy c.beforesalary if age>26
```

在命令窗口输入命令并按回车键进行确认，结果如图 6.14 所示。

```
. anova  nowsalary identity policy c.beforesalary if age>26

                           Number of obs =       25     R-squared     =  0.8985
                           Root MSE      = .541736     Adj R-squared =  0.8782

                  Source |  Partial SS    df       MS           F     Prob > F
              -----------+----------------------------------------------------
                   Model |  51.9704348     4  12.9926087      44.27     0.0000
                         |
                identity |  1.81439507     2  .907197534       3.09     0.0676
                  policy |  .452084267     1  .452084267       1.54     0.2289
            beforesal~y  |  34.8433685     1  34.8433685     118.73     0.0000
                         |
                Residual |  5.86956522    20  .293478261
              -----------+----------------------------------------------------
                   Total |       57.84    24        2.41
```

图 6.14　分析结果图

通过观察分析结果，我们可以看出共有 25 个有效样本参与了方差分析。

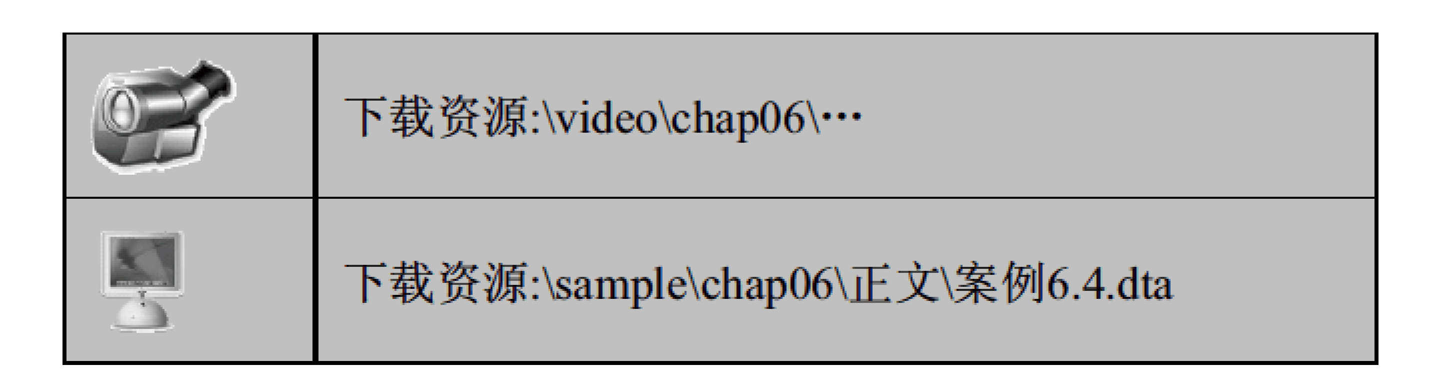
- 可决系数（R-squared）以及修正的可决系数（Adj R-squared）都超过了 80%，说明模型的拟合程度很高，也就是说模型的解释能力很强。
- Prob > F Model=0.0000，说明模型的整体是很显著的。
- Prob > F identity =0.0676，说明变量 identity 的主效应是比较不显著的。
- Prob > F policy =0.2289，说明变量 policy 的主效应也是非常不显著的。
- Prob > F beforesalary =0.0000，说明变量 beforesalary 的主效应是非常显著的。

6.4　实例四——重复测量方差分析

6.4.1　重复测量方差分析的功能与意义

在研究中，我们经常需要对同一个观察对象重复进行多次观测，这样得到的数据称为重复测量资料；而对于重复测量资料进行方差分析就需要采用重复测量方差分析方法。重复测量方差分析与前述的方差分析的最大差别在于：它可以考察测量指标是否会随着测量次数的增加而变化，以及是否会受时间的影响。

6.4.2　相关数据来源



	下载资源:\video\chap06\···
	下载资源:\sample\chap06\正文\案例6.4.dta

【例 6.4】某食品公司为计划改进一种食品的销售策略而提出了一种方案，并随机选择了 20 个销售网点施行销售策略。表 6.4 为所调查网点的实施策略后的一个月的销售量(单位:kg)。通过分析说明这种方案是否有效。

表 6.4　各网点销售量统计表

网点	方案	销售量
1	实施前	70
2	实施前	48
3	实施前	34
4	实施前	56
5	实施前	36
…	…	…
19	实施后	79
20	实施后	67

6.4.3　Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有 3 个变量，分别为网点、方案和销售量。我们把网点变量设定为 number，把方案变量设定为 plan，并且把实施前设定为 1、把实施后设定为 2，把销售量变量设定为 sale，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作我们在第 1 章中已有详细讲述。录入完成后数据如图 6.15 所示。

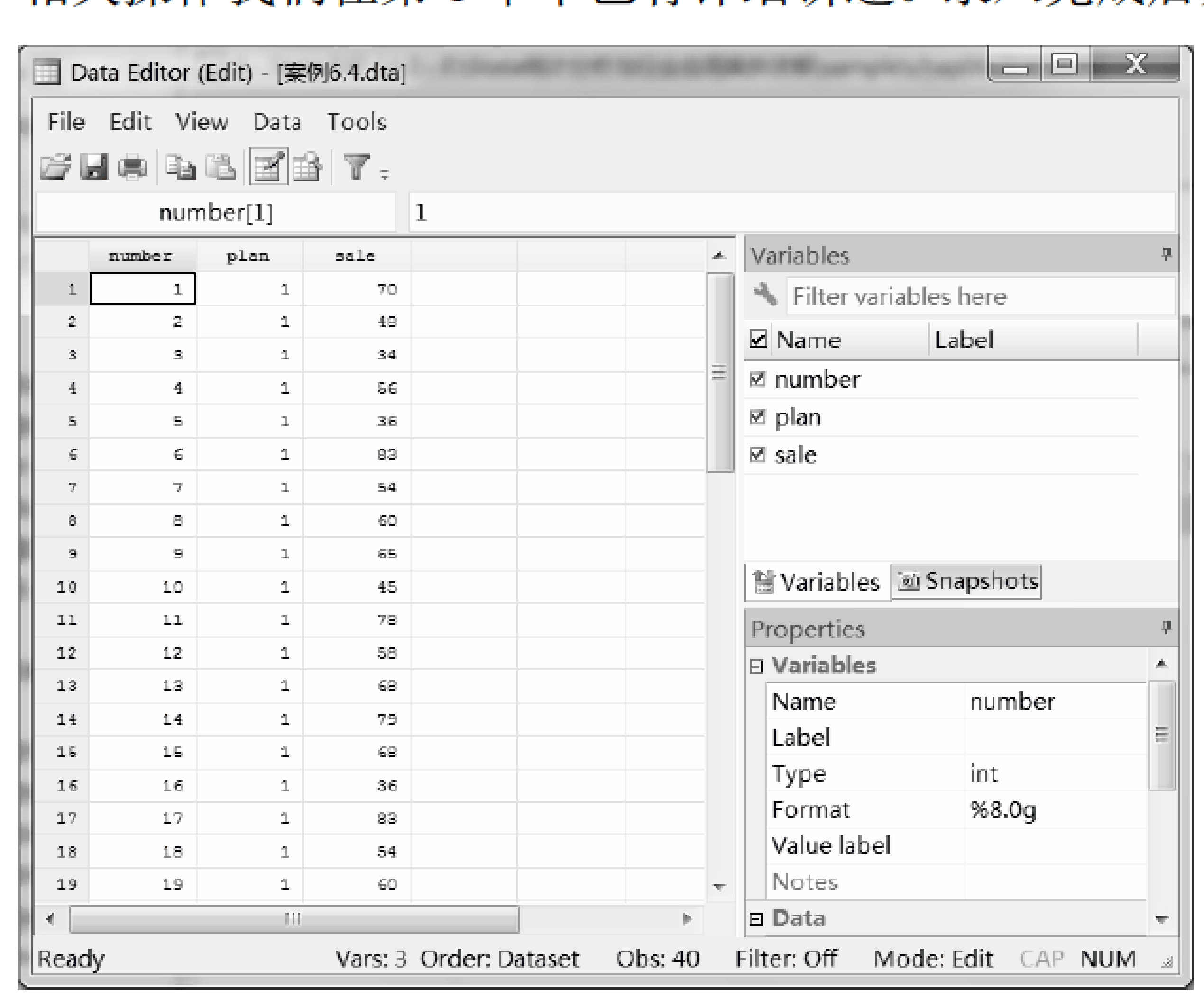


图 6.15　案例 6.4 数据

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入如下命令（旨在分析说明这种方案是否有效）：

```
anova  sale  number plan,repeated(plan)
```

03 设置完毕后，按键盘上的回车键，等待输出结果。

6.4.4 结果分析

我们可以在 Stata 14.0 主界面的结果窗口看到如图 6.16 所示的分析结果。

```
. anova  sale  number plan,repeated(plan)

                          Number of obs =         40    R-squared     =  0.7726
                          Root MSE      =    13.1535    Adj R-squared =  0.5331

                  Source | Partial SS    df       MS          F    Prob > F
              -----------+----------------------------------------------------
                   Model |    11165.5    20    558.275       3.23    0.0067
                         |
                  number |   3241.275    19  170.593421      0.99    0.5121
                    plan |   7924.225     1    7924.225     45.80    0.0000
                         |
                Residual |   3287.275    19  173.014474
              -----------+----------------------------------------------------
                   Total |  14452.775    39  370.583974


Between-subjects error term:  number
                     Levels:  20        (19 df)
     Lowest b.s.e. variable:  number

Repeated variable: plan
                                          Huynh-Feldt epsilon        =  1.0000
                                          Greenhouse-Geisser epsilon =  1.0000
                                          Box's conservative epsilon =  1.0000

                                            ------------ Prob > F ------------
                  Source |     df      F    Regular    H-F      G-G      Box
              -----------+----------------------------------------------------
                    plan |      1    45.80   0.0000   0.0000   0.0000   0.0000
                Residual |     19
              -----------+----------------------------------------------------
```

图 6.16　分析结果图

通过观察分析结果，我们可以看出共有 40 个有效样本参与了方差分析。

- 可决系数（R-squared）以及修正的可决系数（Adj R-squared）都在 50%以上，说明模型的拟合程度还是可以的，也就是说模型的解释能力还是可以的。
- Prob > F Model=0.0067，说明模型的整体是很显著的。
- Prob > F number =0.5121，说明变量 number 的效应是非常不显著的。
- Prob > F plan =0.0000，说明变量 plan 的主效应是非常显著的。

也就是说，销售量的大小与网点是没有太大关系的，网点的差异对销售量差异的影响程度是很不显著的。而方案的实施却对销售量的大小有显著影响。

6.4.5 案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

例如，我们只针对 number 变量大于 3 的观测样本进行重复测量方差分析，那么操作命令即为：

```
anova  sale  number plan if number>3,repeated(plan)
```

在命令窗口输入命令并按回车键进行确认，结果如图 6.17 所示。

```
. anova  sale  number plan if number>3,repeated(plan)

                           Number of obs =      34     R-squared     =  0.7672
                           Root MSE      = 12.5119     Adj R-squared =  0.5199

                  Source |  Partial SS    df       MS           F     Prob > F
              -----------+----------------------------------------------------
                   Model |  8256.29412    17   485.66436       3.10     0.0142
                         |
                  number |  2818.05882    16  176.128676       1.13     0.4083
                    plan |  5438.23529     1  5438.23529      34.74     0.0000
                         |
                Residual |  2504.76471    16  156.547794
              -----------+----------------------------------------------------
                   Total |  10761.0588    33  326.092692


Between-subjects error term:  number
                     Levels:  17        (16 df)
     Lowest b.s.e. variable:  number

Repeated variable: plan
                                          Huynh-Feldt epsilon        =  1.0000
                                          Greenhouse-Geisser epsilon =  1.0000
                                          Box's conservative epsilon =  1.0000

                                            ------------ Prob > F ------------
                  Source |     df      F    Regular    H-F      G-G      Box
              -----------+----------------------------------------------------
                    plan |      1    34.74   0.0000   0.0000   0.0000   0.0000
                Residual |     16
              -----------+----------------------------------------------------
```

图 6.17　分析结果图

通过观察分析结果，我们可以看出共有 34 个有效样本参与了方差分析。

- Prob>F Model= 0.0142，说明模型的整体是很显著的。
- Prob > F number =0.4083，说明变量 number 的效应是非常不显著的。
- Prob > F plan =0.0000，说明变量 plan 的主效应是非常显著的。

6.5　本章习题

（1）表 6.5 给出了 4 种包装对某饮料销售水平影响的测量结果，数据为各大超市 20 天的每日总销售量。试用单因素方差分析检验 4 种包装对饮料销售水平的影响是否相同。

表 6.5　4 种包装下的饮料销售水平

测量编号	总销售量/瓶	包装类别
1	90	1
2	94	1
3	88	1
4	110	1
5	96	1
6	84	2
…	…	…
16	88	4
17	90	4
18	73	4
19	88	4
20	86	4

（2）表 6.6 给出了两种包装和两种口味对某饮料销售水平的影响测量结果，数据为 4 种饮料在 20 家超市一天的总销售量。试用多因素方差分析检验不同包装及口味对饮料销售水平的影响是否相同。

表 6.6　4 种饮料在 20 家超市一天的总销售量

超市编号	销售数量/瓶	包装	口味
1	10	a1	b1
2	10	a1	b1
3	40	a1	b1
4	50	a1	b1
5	10	a1	b1
6	30	a1	b2
…	…	…	…
18	70	a2	b2
19	60	a2	b2
20	30	a2	b2

（3）某医院实施新政策以改善部分年轻医生的生活水平。政策实施后开始对年轻医生待遇的改善情况进行调查，调查结果如表 6.7 所示。用实施新政策后的工资来反映生活水平的提高，要求剔除实施新政策前的工资差异，试分析医生的级别和新政策对年轻医生工资的提高是否有显著的影响。

表 6.7　年轻医生工资表（单位：千元）

年龄	原工资	现工资	医生级别	政策实施
27	4	4	2	否
26	2	5	3	否
26	3	4	1	是
28	3	5	2	否
29	4	5	2	是
…	…	…	…	…
29	6	9	3	是
27	8	10	2	否

（4）某建材公司为计划改进一种钢管的销售策略而提出了一种方案，并随机选择了 20 个销售网点，施行不同的销售策略。表 6.8 为所调查网点实施策略后的一个月的销售量（单位：个）。通过分析说明这种方案是否有效。

表 6.8　各网点销售量统计表

网点	方案	销售量
1	实施前	56
2	实施前	36
3	实施前	34
4	实施前	79
5	实施前	67
…	…	…
19	实施后	28
20	实施后	45

第 7 章　Stata 相关分析

在得到相关数据资料后，我们要对这些数据进行分析，研究各个变量之间的关系。相关分析是应用非常广泛的一种方法。它是不考虑变量之间的因果关系而只研究分析变量之间的相关关系的一种统计分析方法，常用的相关分析包括简单相关分析、偏相关分析等。下面我们将分别介绍这些方法在实例中的应用。

7.1　实例一——简单相关分析

7.1.1　简单相关分析的功能与意义

Stata 的简单相关分析（Bivariate）是最简单也是最常用的一种相关分析（Correlate）方法，其基本功能是可以研究变量间的线性相关程度并用适当的统计指标表示出来。

7.1.2　相关数据来源

	下载资源:\video\chap07\…
	下载资源:\sample\chap07\正文\案例7.1.dta

【例 7.1】表 7.1 给出了杭州市 2006 年市区分月统计的平均温度和日照时数。试据此分析平均温度和日照时数的相关性。

表 7.1　杭州市 2006 年市区分月部分气象概况统计

月份	平均温度/℃	日照时数/h
1	5.8	62.1
2	6.2	58.6
3	12.5	137.9
4	18.3	154.8
5	21.5	131.4
6	25.9	119.5
7	30.1	183.8
8	30.6	215.6
9	23.3	96.9
10	21.9	91.9
11	15.2	81.3
12	7.7	89

7.1.3 Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有 3 个变量，分别是月份、平均温度和日照时数。我们把月份变量设定为 month，把平均温度变量设定为 tem，把日照时数变量设定为 hour，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作我们在第 1 章中已有详细讲述。录入完成后数据如图 7.1 所示。

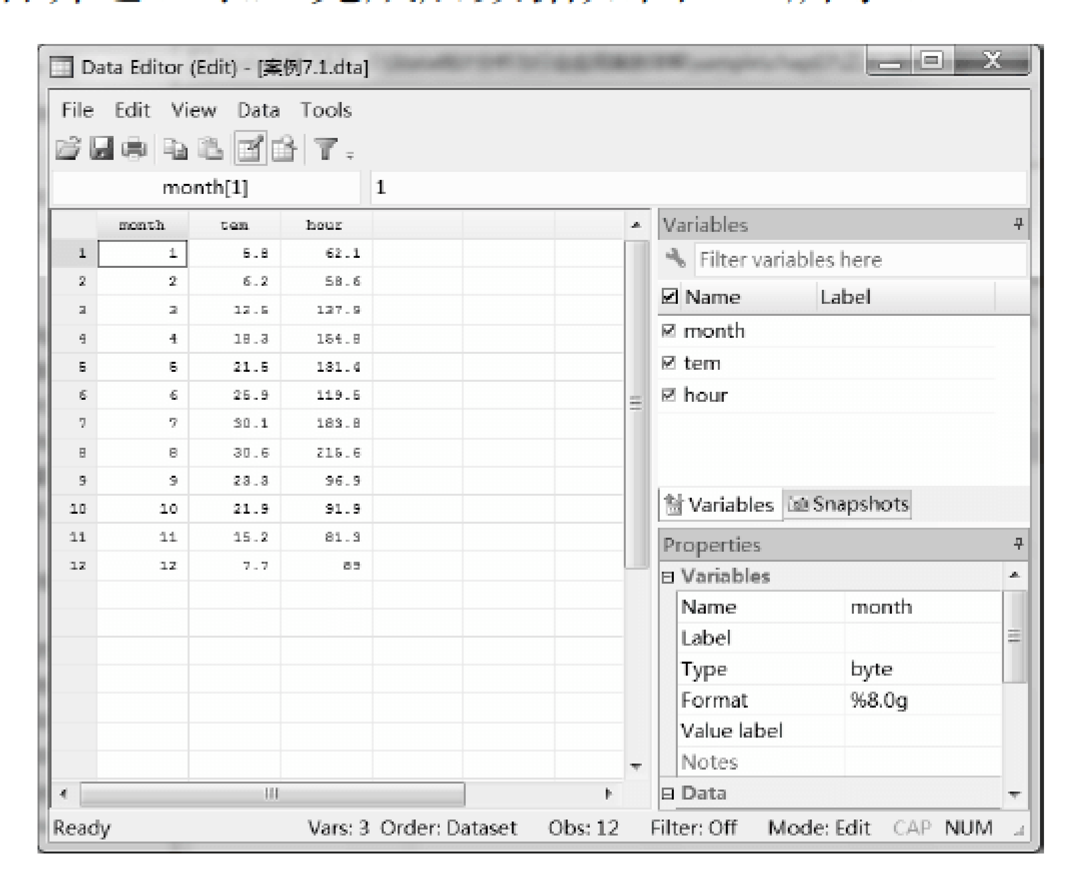


图 7.1　案例 7.1 数据

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

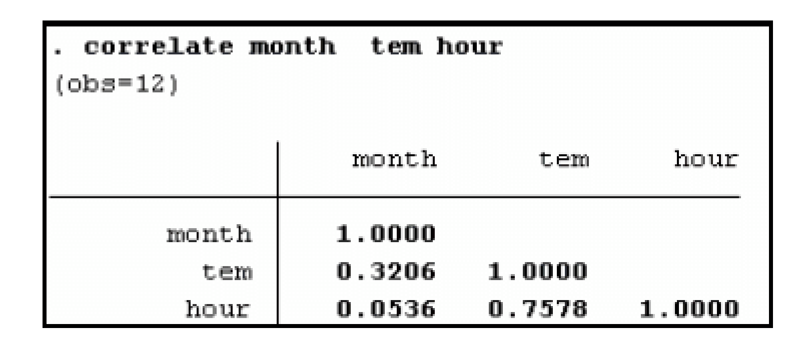
02 在主界面的“Command”文本框中输入命令（对月份、平均温度和日照时数 3 个变量进行简单相关分析）：

```
correlate month  tem hour
```

03 设置完毕后，按键盘上的回车键，等待输出结果。

7.1.4 结果分析

我们可以在 Stata 14.0 主界面的结果窗口看到如图 7.2 所示的分析结果。



. correlate month tem hour
(obs=12)

	month	tem	hour
month	1.0000		
tem	0.3206	1.0000	
hour	0.0536	0.7578	1.0000

图 7.2　分析结果图

从上述分析结果中可以得到很多信息。首先可以看到共有 12 个样本参与了分析(obs=12)，然后可以看到变量两两之间的相关系数，其中 month 与 tem 之间的相关系数是 0.3206，month 与 hour 之间的相关系数是 0.0536，tem 与 hour 之间的相关系数是 0.7578，所以本例的结论是平均温度和日照时数具有比较高的正相关性。

7.1.5 案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

1．延伸 1：获得变量的方差协方差矩阵

我们在进行数据分析时，很多时候需要使用变量的方差协方差矩阵。该操作对应的 Stata 命令是：

```
correlate month  tem hour,covariance
```

在命令窗口输入命令并按回车键进行确认，结果如图 7.3 所示。

```
. correlate month  tem hour,covariance
(obs=12)

             |    month      tem     hour
-------------+---------------------------
       month |       13
         tem |  10.1909  77.7027
        hour |  9.34546  323.211  2341.01
```

图 7.3　分析结果图

从上述分析结果中可以看到变量的方差协方差矩阵，其中 month 的方差是 13，tem 的方差是 77.7027，hour 的方差是 2341.01，month 与 tem 的协方差是 10.1909，month 与 hour 的协方差是 9.34546，tem 与 hour 之间的相关系数是 323.211。

2．延伸 2：获得相关性的显著性检验

该操作对应的 Stata 命令是：

```
pwcorr month  tem hour,sig
```

在命令窗口输入命令并按回车键进行确认，结果如图 7.4 所示。

```
. pwcorr month  tem hour,sig

             |    month      tem     hour
-------------+---------------------------
       month |   1.0000
             |
             |
         tem |   0.3206   1.0000
             |   0.3096
             |
        hour |   0.0536   0.7578   1.0000
             |   0.8687   0.0043
             |
```

图 7.4　分析结果图

从上述分析结果中可以看到变量的相关性的显著性检验结果。其中，month 与 tem 之间的相关性显著性 P 值是 0.3096，month 与 hour 之间的相关性显著性 P 值是 0.8687，hour 与 tem 之间的相关性显著性 P 值是 0.0043。

此外，还有一种更为精确的 sidak 方法。该操作对应的 Stata 命令是：

```
pwcorr month  tem hour,sidak sig
```

在命令窗口输入命令并按回车键进行确认，结果如图 7.5 所示。

```
. pwcorr month  tem hour,sidak sig

             |    month      tem     hour
-------------+---------------------------
       month |   1.0000
             |
             |
         tem |   0.3206   1.0000
             |   0.6709
             |
        hour |   0.0536   0.7578   1.0000
             |   0.9977   0.0128
```

图 7.5　分析结果图

从上述分析结果中可以看到变量的相关性的显著性检验结果。其中，month 与 tem 之间的相关性显著性 P 值是 0.6709，month 与 hour 之间的相关性显著性 P 值是 0.9977，hour 与 tem 之间的相关性显著性 P 值是 0.0128。

3．延伸 3：获得相关性的显著性检验，并进行标注

很多时候我们希望能够一目了然地看出变量相关在不同的置信水平上是否显著，例如置信水平为 99%时，对应的 Stata 命令是：

```
pwcorr month  tem hour,sidak sig star(0.01)
```

在命令窗口输入命令并按回车键进行确认，结果如图 7.6 所示。

```
. pwcorr month  tem hour,sidak sig star(0.01)

             |    month      tem     hour
-------------+---------------------------
       month |   1.0000
             |
             |
         tem |   0.3206   1.0000
             |   0.6709
             |
        hour |   0.0536   0.7578   1.0000
             |   0.9977   0.0128
```

图 7.6　分析结果图

从上述分析结果图中可以看出所有变量间的相关关系不显著。如果把置信水平换成 90%，那么对应的 Stata 命令是：

```
pwcorr month  tem hour,sidak sig star(0.10)
```

在命令窗口输入命令并按回车键进行确认，结果如图 7.7 所示。

```
. pwcorr month  tem hour,sidak sig star(0.10)

             |    month      tem     hour
-------------+---------------------------
       month |   1.0000
             |
             |
         tem |   0.3206   1.0000
             |   0.6709
             |
        hour |   0.0536   0.7578*  1.0000
             |   0.9977   0.0128
```

图 7.7　分析结果图

可以看出在 90%的置信水平下，仅有 hour 与 tem 的相关性是显著的。

7.2　实例二——偏相关分析

7.2.1　偏相关分析的功能与意义

很多情况下，需要进行相关分析的变量的取值会同时受到其他变量的影响，这时就需要把其他变量控制住，然后输出控制其他变量影响后的相关系数。Stata 的偏相关分析（Partial）过程就是为解决这一问题而设计的。

7.2.2　相关数据来源

	下载资源:\video\chap07\…
	下载资源:\sample\chap07\正文\案例7.2.dta

【例 7.2】表 7.2 给出了随机抽取的山东省某学校的 12 名学生的 IQ 值、语文成绩和数学成绩。因为语文成绩和数学成绩都受 IQ 的影响，所以试用偏相关分析研究学生语文成绩和数学成绩的相关关系。

表 7.2　12 名学生的 IQ、语文成绩和数学成绩

IQ	语文成绩	数学成绩
100	86	85
120	93	98
117	91	90
98	82	79
60	43	32
62	45	37
88	60	61
123	99	98
110	88	89
115	86	91
116	90	91
71	67	63

7.2.3　Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有 3 个变量，分别是 IQ、语文成绩和数学成绩。我们把 IQ 变量设定为 IQ，把语文成绩变量设定为 YW，把数学成绩变量设定为 SX，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作我们在第 1

章中已有详细讲述。录入完成后数据如图 7.8 所示。

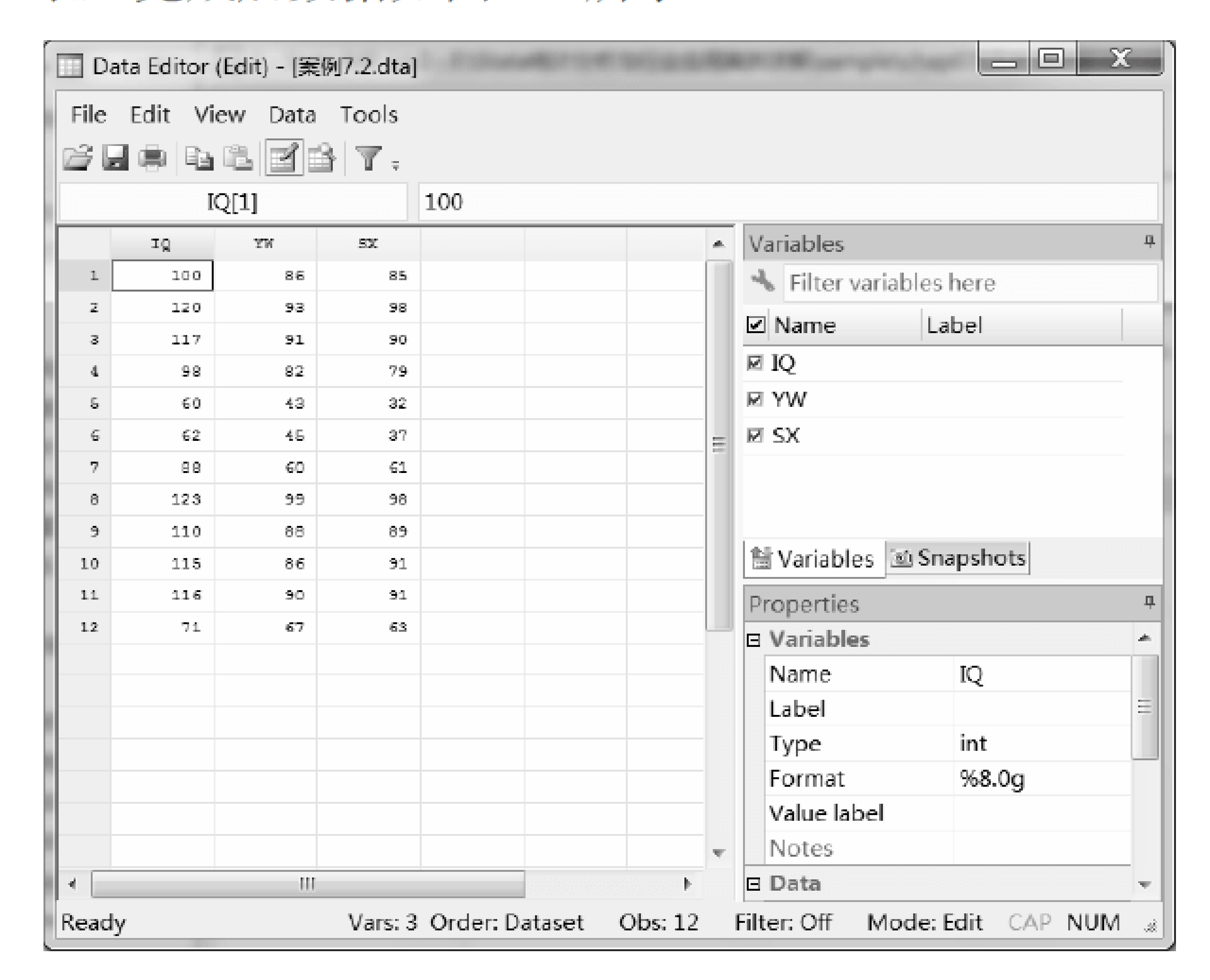


图 7.8　案例 7.2 数据

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

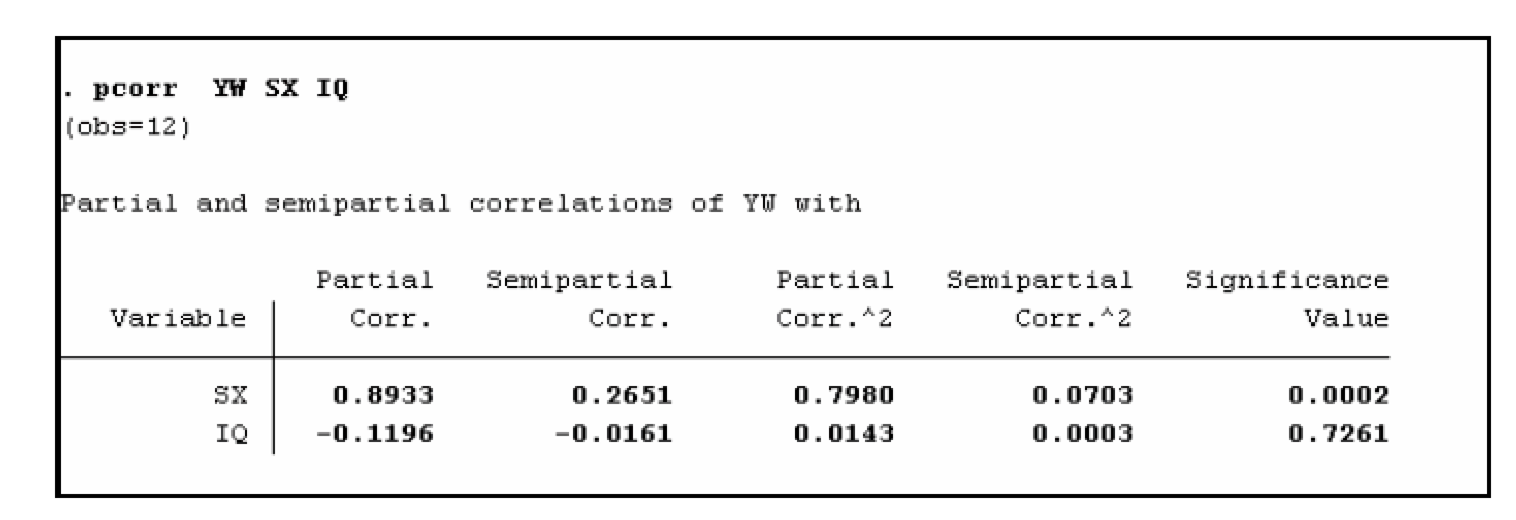
02 在主界面的“Command”文本框中输入命令：

```
pcorr  YW SX IQ
```

03 设置完毕后，按键盘上的回车键，等待输出结果。

7.2.4　结果分析

在 Stata 14.0 主界面的结果窗口我们可以看到如图 7.9 所示的分析结果。



```
. pcorr  YW SX IQ
(obs=12)

Partial and semipartial correlations of YW with

               Partial  Semipartial   Partial  Semipartial  Significance
   Variable      Corr.        Corr.   Corr.^2      Corr.^2         Value
-------------+----------------------------------------------------------
         SX     0.8933       0.2651    0.7980       0.0703        0.0002
         IQ    -0.1196      -0.0161    0.0143       0.0003        0.7261
```

图 7.9　分析结果图

通过观察分析结果，我们可以看出共有 12 个有效样本参与了方差分析，在控制住 IQ 变量的情况下，语文成绩和数学成绩的偏相关系数（Partial Corr.）是 0.8933，显著性水平（Significance Value）是 0.0002。此外，该结果还给出了控制住数学成绩变量的情况下，语文成绩和 IQ 之间的偏相关关系，它们的偏相关系数（Partial Corr.）是-0.1196，显著性水平（Significance Value）是 0.7261。

7.2.5 案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

例如，我们仅用偏相关分析研究 IQ 值在 100 以上的学生语文成绩和数学成绩的相关关系。该操作对应的 Stata 命令是：

```
pcorr  YW SX IQ if  IQ>100
```

在命令窗口输入命令并按回车键进行确认，结果如图 7.10 所示。

```
. pcorr  YW SX IQ if  IQ>100
(obs=6)

Partial and semipartial correlations of YW with

               Partial   Semipartial      Partial   Semipartial   Significance
   Variable      Corr.         Corr.      Corr.^2       Corr.^2          Value

         SX     0.2312        0.1200       0.0535        0.0144         0.7082
         IQ     0.5291        0.3149       0.2800        0.0992         0.3592
```

图 7.10　分析结果图

通过观察分析结果，我们可以看出共有 6 个有效样本参与了方差分析，在控制住 IQ 变量的情况下，语文成绩和数学成绩的偏相关系数(Partial Corr.)是 0.2312，显著性水平(Significance Value）是 0.7082。此外，该结果还给出了控制住数学成绩变量的情况下，语文成绩和 IQ 之间的偏相关关系，它们的偏相关系数（Partial Corr.）是 0.5291，显著性水平（Significance Value）是 0.3592。

7.3 本章习题

（1）表 7.3 给出了铁岭、朝阳和葫芦岛 2006 年各月的平均气温情况。试用简单相关分析方法研究这 3 个地区月平均气温的相关性。

表 7.3　铁岭、朝阳、葫芦岛 2006 年各月平均气温统计（单位：℃）

月份	铁岭	朝阳	葫芦岛
1	−12.3	−8.1	−7.0
2	−8.2	−5.8	−4.3
3	0.8	3.0	2.8
4	7.6	9.4	9.3
5	18.3	19.2	18.3
6	21.3	23.3	21.5
7	24.2	24.5	24.3
8	23.9	24.5	24.3
9	17.9	18.1	20.3
10	11.6	12.1	13.8
11	0.4	1.2	3.1
12	−6.5	−6.5	−3.8

（2）某研究者对当地的塑料制品厂的工人工龄、性别、年龄和月工资等情况展开了调查，数据如表 7.4 所示。

表 7.4　某塑料制品厂的工人情况表

编号	工龄/年	性别	年龄	月工资/元
001	1	男	20	700
002	1	男	21	700
…	…	…	…	…
104	2	女	22	800
105	2	女	21	1000
106	2	女	20	900

① 试在控制住性别变量的情况下研究年龄与月工资的偏相关关系。
② 试在控制住工龄变量的情况下研究年龄与月工资的偏相关关系。
③ 试在控制住年龄变量的情况下研究工龄与月工资的偏相关关系。

第 8 章　Stata 主成分分析与因子分析

在进行数据统计分析时，还往往会遇见变量特别多的情况，而且很多时候这些变量之间还存在着很强的相关关系或者说变量之间存在着很强的信息重叠，如果我们直接对数据进行分析，一方面会带来工作量的无谓加大，另一方面还会出现一些模型应用的错误，于是主成分分析与因子分析应运而生。这两种分析方法的基本思想都是在不损失大量信息的前提下，利用较少的独立变量来替代原来的变量进行进一步的分析。下面我们将分别介绍这两种方法在实例中的应用。

8.1　实例一——主成分分析

8.1.1　主成分分析的功能与意义

在实际工作中，往往会出现所搜集的变量间存在较强相关关系的情况。如果直接利用数据进行分析，不仅会使模型变得很复杂，还会带来多重共线性等问题。主成分分析提供了解决这一问题的方法，其基本思想是将众多的初始变量整合成少数几个互相无关的主成分变量，而这些新的变量尽可能地包含了初始变量的全部信息，然后利用这些新的变量来替代以前的变量进行分析。

8.1.2　相关数据来源

	下载资源:\video\chap08\…
	下载资源:\sample\chap08\正文\案例8.1.dta

【例 8.1】表 8.1 给出了我国近年来国民经济的主要指标统计（1998－2005）。试用主成分分析法对这些指标提取主成分并写出提取的主成分与这些指标之间的表达式。

表 8.1　我国近年来国民经济的主要指标统计（1998－2005）

年份	全国人口/万人	农林牧渔业总产值/亿元	…	粮食/万吨	棉花/万吨	油料/万吨
1998	124 810.0	24 516.7	…	51 230.0	450.1	2 313.9
1999	125 909.0	24 519.1	…	50 839.0	382.9	2 601.2
2000	126 743.0	24 915.8	…	46 218.0	442.0	2 955.0
2001	127 627.0	26 179.6	…	45 264.0	532.4	2 864.9
2002	128 453.0	27 390.8	…	45 706.0	491.6	2 897.2

（续表）

年份	全国人口/万人	农林牧渔业总产值/亿元	…	粮食/万吨	棉花/万吨	油料/万吨
2003	129 227.0	29 691.8	…	43 069.5	486.0	2 811.0
2004	229 988.0	36 239.0	…	46 946.9	632.4	3 065.9
2005	130 756.0	39 450.9	…	48 402.2	571.4	3 077.1

8.1.3 Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有 19 个变量，分别是年份、全国人口（万人）、农林牧渔业总产值（亿元）、工业总产值（亿元）、国内生产总值（亿元）、全社会投资总额（亿元）、货物周转量（亿吨千米）、社会消费品零售总额（亿元）、进出口贸易总额（亿元）、原煤（亿吨）、发电量（亿千瓦时）、原油（万吨）、钢（万吨）、汽车（万辆）、布（亿米）、糖（万吨）、粮食（万吨）、棉花（万吨）和油料（万吨）。我们把这些变量分别定义为 V1、V2、V3、V4、V5、V6、V7、V8、V9、V10、V11、V12、V13、V14、V15、V16、V17、V18、V19。变量类型及长度采取系统默认方式，然后录入相关数据。相关操作我们在第 1 章中已有详细讲述。录入完成后数据如图 8.1 所示。

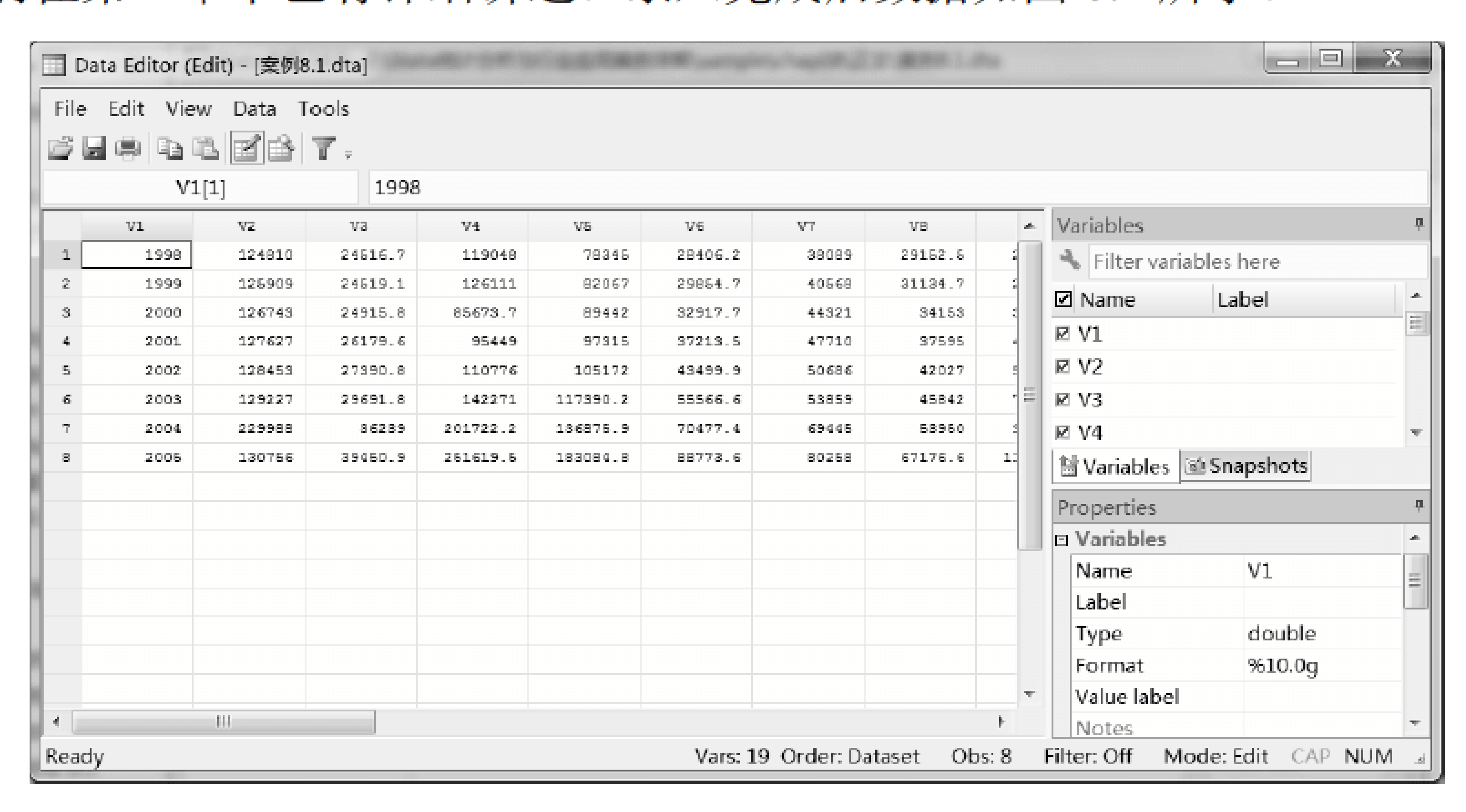


图 8.1　案例 8.1 数据

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中分别输入如下命令并按键盘上的回车键进行确认。

- correlate V2-V19：本命令的含义是对全国人口（万人）、农林牧渔业总产值（亿元）、工业总产值（亿元）、国内生产总值（亿元）、全社会投资总额（亿元）、货物周转量（亿吨千米）、社会消费品零售总额（亿元）、进出口贸易总额（亿元）、原煤（亿吨）、发电量（亿千瓦时）、原油（万吨）、钢（万吨）、汽车（万辆）、布（亿米）、糖（万吨）、粮食（万吨）、棉花（万吨）和油料（万吨）等变量进行相关性分析。
- pca V2-V19：本命令的含义是对全国人口（万人）、农林牧渔业总产值（亿元）、工业

总产值（亿元）、国内生产总值（亿元）、全社会投资总额（亿元）、货物周转量（亿吨千米）、社会消费品零售总额（亿元）、进出口贸易总额（亿元）、原煤（亿吨）、发电量（亿千瓦时）、原油（万吨）、钢（万吨）、汽车（万辆）、布（亿米）、糖（万吨）、粮食（万吨）、棉花（万吨）和油料（万吨）等变量进行主成分分析。

03 设置完毕后，按键盘上的回车键，等待输出结果。

8.1.4 结果分析

在 Stata 14.0 主界面的结果窗口可以看到如图 8.2~图 8.4 所示的分析结果。

```
. correlate V2-V19
(obs=8)
```

	V2	V3	V4	V5	V6	V7	V8
V2	1.0000						
V3	0.5412	1.0000					
V4	0.4583	0.9489	1.0000				
V5	0.3417	0.9720	0.9144	1.0000			
V6	0.4542	0.9907	0.9299	0.9880	1.0000		
V7	0.4922	0.9907	0.9119	0.9849	0.9903	1.0000	
V8	0.4012	0.9772	0.8980	0.9942	0.9932	0.9916	1.0000
V9	0.4943	0.9899	0.9141	0.9798	0.9975	0.9910	0.9898
V10	0.4934	0.9698	0.9386	0.9390	0.9715	0.9428	0.9476
V11	0.4689	0.9752	0.8829	0.9743	0.9920	0.9832	0.9910
V12	0.4458	0.9850	0.9047	0.9874	0.9954	0.9906	0.9949
V13	0.4223	0.9854	0.9274	0.9907	0.9985	0.9875	0.9954
V14	0.4968	0.9539	0.8654	0.9396	0.9748	0.9519	0.9647
V15	0.4688	0.9855	0.9002	0.9852	0.9959	0.9942	0.9963
V16	0.4220	0.5298	0.5585	0.4544	0.5537	0.4680	0.5010
V17	-0.0671	-0.1303	0.1409	-0.1855	-0.2125	-0.2022	-0.2534
V18	0.7144	0.8243	0.6608	0.7470	0.7887	0.8251	0.7852
V19	0.4225	0.6745	0.4507	0.7018	0.6941	0.7535	0.7403

	V9	V10	V11	V12	V13	V14	V15
V9	1.0000						
V10	0.9642	1.0000					
V11	0.9943	0.9588	1.0000				
V12	0.9944	0.9652	0.9913	1.0000			
V13	0.9933	0.9705	0.9915	0.9944	1.0000		
V14	0.9795	0.9626	0.9898	0.9708	0.9737	1.0000	
V15	0.9972	0.9569	0.9961	0.9974	0.9946	0.9766	1.0000
V16	0.5551	0.6643	0.5781	0.5264	0.5581	0.6818	0.5337
V17	-0.2490	-0.1469	-0.3237	-0.2436	-0.2129	-0.3604	-0.2651
V18	0.8058	0.7884	0.8085	0.8092	0.7752	0.7842	0.8098
V19	0.7278	0.5399	0.7457	0.7180	0.6867	0.7043	0.7488

	V16	V17	V18	V19
V16	1.0000			
V17	-0.2006	1.0000		
V18	0.3122	-0.3299	1.0000	
V19	0.1570	-0.5760	0.6735	1.0000

图 8.2 分析结果图

图 8.2 展示的是参与主成分分析的所有变量之间的方差-协方差矩阵。关于本命令以及结果我们在前面章节中已经介绍过，此处不再赘述。可以发现，本例中有很多变量之间的相关关系是非常强的，有些甚至超过了 90%，这说明变量之间存在着相当数量的信息重叠。我们进行主成分分析把众多的初始变量整合成少数几个互相之间无关的主成分变量是非常有必要的。

图 8.3 展示的是主成分分析的结果。其中最左列（Component）表示的是系统提取的主成分名称，可以发现，我们的 Stata 总共提取了 18 个主成分。Eigenvalue 列表示的是系统提取的主成分的特征值，特征值的大小意味着该主成分的解释能力，特征值越大解释能力越强，可以发现 Stata 提取的 18 个主成分中只有前 7 个是有效的，因为 Comp8~Comp18 的特征值（Eigenvalue）均为 0。Proportion 列表示的是系统提取的主成分的方差贡献率，方差贡献率同样表示主成分的解释能力，可以发现第 1 个主成分的方差贡献率为 0.8023，表示该主成分解释

了所有变量 80.23%的信息。第 2 个主成分的方差贡献率为 0.0788，表示该主成分解释了所有变量 7.88%的信息，依次类推。Cumulative 列表示的是主成分的累计方差贡献率，其中前两个主成分的方差贡献率为 0.8812，前 3 个主成分的方差贡献率为 0.9362，依次类推。

```
. pca V2-V19

Principal components/correlation                  Number of obs    =          8
                                                  Number of comp.  =          7
                                                  Trace            =         18
    Rotation: (unrotated = principal)             Rho              =     1.0000
```

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	14.442	13.0228	0.8023	0.8023
Comp2	1.41918	.429462	0.0788	0.8812
Comp3	.989717	.118447	0.0550	0.9362
Comp4	.87127	.629391	0.0484	0.9846
Comp5	.241878	.214668	0.0134	0.9980
Comp6	.0272104	.0184781	0.0015	0.9995
Comp7	.00873232	.00873232	0.0005	1.0000
Comp8	0	0	0.0000	1.0000
Comp9	0	0	0.0000	1.0000
Comp10	0	0	0.0000	1.0000
Comp11	0	0	0.0000	1.0000
Comp12	0	0	0.0000	1.0000
Comp13	0	0	0.0000	1.0000
Comp14	0	0	0.0000	1.0000
Comp15	0	0	0.0000	1.0000
Comp16	0	0	0.0000	1.0000
Comp17	0	0	0.0000	1.0000
Comp18	0	.	0.0000	1.0000

图 8.3　分析结果图

Principal components (eigenvectors)

Variable	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6	Comp7	Unexplained
V2	0.1377	-0.0208	0.7802	0.3558	0.2120	-0.2517	-0.1105	0
V3	0.2605	0.0925	-0.0038	0.0889	0.0016	-0.1416	-0.0631	0
V4	0.2390	0.3401	-0.0339	0.0407	0.0747	-0.4336	0.4109	0
V5	0.2560	0.0458	-0.2246	0.0239	-0.0020	0.0019	0.0416	0
V6	0.2618	0.0460	-0.0787	-0.0174	-0.0045	-0.1802	-0.0416	0
V7	0.2606	0.0096	-0.0818	0.1116	0.0627	0.0709	0.1326	0
V8	0.2600	-0.0069	-0.1492	-0.0001	0.0119	0.1970	0.2038	0
V9	0.2625	0.0076	-0.0403	-0.0028	0.0441	-0.2720	-0.2868	0
V10	0.2550	0.1459	0.0493	-0.1095	-0.2645	0.0447	-0.3505	0
V11	0.2620	-0.0452	-0.0439	-0.0632	-0.0008	-0.0218	0.1900	0
V12	0.2614	0.0088	-0.0923	0.0028	-0.0558	0.1706	-0.5951	0
V13	0.2610	0.0499	-0.1051	-0.0420	-0.0156	0.0143	0.2206	0
V14	0.2587	-0.0437	0.0466	-0.1801	0.0040	-0.0858	0.0383	0
V15	0.2623	-0.0147	-0.0731	0.0053	0.0374	0.0878	-0.1425	0
V16	0.1504	0.1645	0.4389	-0.7042	0.1901	0.3600	0.1158	0
V17	-0.0679	0.7491	-0.0427	0.3658	0.2422	0.4021	-0.0445	0
V18	0.2187	-0.1718	0.2611	0.3456	-0.6133	0.3933	0.2633	0
V19	0.1913	-0.4745	-0.0939	0.2263	0.6306	0.2924	0.0440	0

图 8.4　分析结果图

图 8.4 展示的是主成分特征向量矩阵，以表明各个主成分在各个变量上的载荷，从而可以得出各主成分的表达式。值得一提的是，在表达式中各个变量已经不是原始变量，而是标准化变量。其中，前两个特征值比较大的主成分的表达式是：

comp1= 0.1377*全国人口+0.2605*农林牧渔业总产值+0.2390*工业总产值+0.2560*国内生产总值+0.2618*全社会投资总额+0.2606*货物周转量+0.2600*社会消费品零售总额+0.2625*进出口贸易总额+0.2550*原煤+0.2620*发电量+0.2614*原油+0.2610*钢+0.2587*汽车+0.2623*布+0.1504*糖-0.0679*粮食+0.2187*棉花+0.1913*油料

comp2=-0.0208*全国人口+0.0925*农林牧渔业总产值+0.3401 *工业总产值+0.0458*国内生产总值+0.0460*全社会投资总额+0.0096*货物周转量-0.0069*社会消费品零售总额+0.0076*进出口贸易总额+0.1459*原煤-0.0452*发电量+0.0088*原油+0.0499*钢-0.0437*汽车-0.0147*布+0.1645*糖+0.7491*粮食-0.1718*棉花-0.4745*油料

在第 1 主成分中，除粮食变量（V17）以外的变量系数比较大，可以看成是反映那些变量的综合指标；在第 2 主成分中，粮食变量的系数比较大，可以看作是反映粮食的综合指标。

因为主成分分析只不过是一种矩阵变换，所以各个主成分并不一定具有实际意义，本例中各个主成分的内在含义就不是很明确。

8.1.5 案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

1．延伸 1：只保留特征值大于 1 的主成分

在上例中可以看到，Stata 总共提取了 7 个有效的主成分，但是只有前两个主成分的特征值是大于 1 的，而且前两个主成分的方差贡献率达到了 0.8812，基本上能够满足我们进行主成分分析的初衷。那么能否只保留特征值大于 1 的主成分呢？

在本节的例子中，操作命令应该相应地修改为：

```
pca V2-V19,mineigen(1)
```

在命令窗口输入命令并按回车键进行确认，结果如图 8.5~图 8.6 所示。

```
. pca V2-V19,mineigen(1)

Principal components/correlation                 Number of obs    =          8
                                                 Number of comp.  =          2
                                                 Trace            =         18
    Rotation: (unrotated = principal)            Rho              =     0.8812
```

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	14.442	13.0228	0.8023	0.8023
Comp2	1.41918	.429462	0.0788	0.8812
Comp3	.989717	.118447	0.0550	0.9362
Comp4	.87127	.629391	0.0484	0.9846
Comp5	.241878	.214668	0.0134	0.9980
Comp6	.0272104	.0184781	0.0015	0.9995
Comp7	.00873232	.00873232	0.0005	1.0000
Comp8	0	0	0.0000	1.0000
Comp9	0	0	0.0000	1.0000
Comp10	0	0	0.0000	1.0000
Comp11	0	0	0.0000	1.0000
Comp12	0	0	0.0000	1.0000
Comp13	0	0	0.0000	1.0000
Comp14	0	0	0.0000	1.0000
Comp15	0	0	0.0000	1.0000
Comp16	0	0	0.0000	1.0000
Comp17	0	0	0.0000	1.0000
Comp18	0	.	0.0000	1.0000

图 8.5 分析结果图

图 8.5 展示的内容与上例一致。

Principal components (eigenvectors)

Variable	Comp1	Comp2	Unexplained
V2	0.1377	-0.0208	.7255
V3	0.2605	0.0925	.007487
V4	0.2390	0.3401	.01052
V5	0.2560	0.0458	.05045
V6	0.2618	0.0460	.007295
V7	0.2606	0.0096	.01872
V8	0.2600	-0.0069	.02349
V9	0.2625	0.0076	.004818
V10	0.2550	0.1459	.03091
V11	0.2620	-0.0452	.005712
V12	0.2614	0.0088	.01307
V13	0.2610	0.0499	.01297
V14	0.2587	-0.0437	.03062
V15	0.2623	-0.0147	.006042
V16	0.1504	0.1645	.635
V17	-0.0679	0.7491	.137
V18	0.2187	-0.1718	.2674
V19	0.1913	-0.4745	.1519

图 8.6　分析结果图

图 8.6 展示的是仅保留特征值大于 1 的主成分的结果，本例中只有前两个主成分的特征值大于 1，所以只保留了前两个主成分进行分析。值得说明的是，图 8.6 最后一列（Unexplained）表示的是该变量未被系统提取的两个主成分解释的信息比例，例如变量 V2 未被解释的信息比例就是 72.55%。这种信息丢失的情况是我们舍弃其他主成分必然付出的代价。

2．延伸 2：限定提取的主成分个数

在有些情况下，可能受某些条件的制约，我们仅能挑选出在规定数目以下的主成分进行分析。那么，我们能否限定提取的主成分的个数呢？

在本节的例子中，例如我们只想提取一个主成分进行分析，那么操作命令应该相应地修改为：

```
pca V2-V19,components(1)
```

在命令窗口输入命令并按回车键进行确认，结果如图 8.7 所示。

. pca V2-V19,components(1)

Principal components/correlation　　　Number of obs = 8
　　　　　　　　　　　　　　　　　　　Number of comp. = 1
　　　　　　　　　　　　　　　　　　　Trace = 18
Rotation: (unrotated = principal)　　　Rho = 0.8023

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	14.442	13.0228	0.8023	0.8023
Comp2	1.41918	.429462	0.0788	0.8812
Comp3	.989717	.118447	0.0550	0.9362
Comp4	.87127	.629391	0.0484	0.9846
Comp5	.241878	.214668	0.0134	0.9980
Comp6	.0272104	.0184781	0.0015	0.9995
Comp7	.00873232	.00873232	0.0005	1.0000
Comp8	0	0	0.0000	1.0000
Comp9	0	0	0.0000	1.0000
Comp10	0	0	0.0000	1.0000
Comp11	0	0	0.0000	1.0000
Comp12	0	0	0.0000	1.0000
Comp13	0	0	0.0000	1.0000
Comp14	0	0	0.0000	1.0000
Comp15	0	0	0.0000	1.0000
Comp16	0	0	0.0000	1.0000
Comp17	0	0	0.0000	1.0000
Comp18	0	.	0.0000	1.0000

（a）

Principal components (eigenvectors)

Variable	Comp1	Unexplained
V2	0.1377	.7261
V3	0.2605	.01963
V4	0.2390	.1747
V5	0.2560	.05343
V6	0.2618	.01029
V7	0.2606	.01885
V8	0.2600	.02356
V9	0.2625	.004899
V10	0.2550	.06112
V11	0.2620	.008606
V12	0.2614	.01318
V13	0.2610	.0165
V14	0.2587	.03333
V15	0.2623	.006349
V16	0.1504	.6734
V17	-0.0679	.9333
V18	0.2187	.3092
V19	0.1913	.4715

（b）

图 8.7　分析结果图

图 8.7（a）展示的内容与上例一致。

图 8.7（b）展示的是我们只提取一个主成分进行分析的结果，该图最后一列（Unexplained）同样说明的是该变量未被系统提取的一个主成分解释的信息比例，例如变量 V2 未被解释的信息比例就是 72.61%。这种信息丢失的情况同样也是我们舍弃其他主成分必然付出的代价。

8.2　实例二——因子分析

8.2.1　因子分析的功能与意义

因子分析在一定程度上可被视作主成分分析的深化和拓展，它对相关问题的研究更为深入透彻。因子分析的基本原理是将具有一定相关关系的多个变量综合为数量较少的几个因子，从而研究一组具有错综复杂关系的实测指标是如何受少数几个内在的独立因子所支配的，所以它属于多元分析中处理降维问题的一种常用的统计方法。

8.2.2　相关数据来源

	下载资源:\video\chap08\…
	下载资源:\sample\chap08\正文\案例8.2.dta

【例 8.2】表 8.2 同样给出了我国近年来国民经济的主要指标统计（1992－2000 年）数据。试用因子分析法对这些指标提取公因子并写出提取的公因子与这些指标之间的表达式。

表 8.2　我国近年来国民经济的主要指标统计（1992－2000 年）

年份	工业总产值/亿元	国内生产总值/亿元	货物周转量/亿吨千米	原煤/亿吨	发电量/亿千瓦时	原油/万吨
1992	37 066.0	26 638.1	29 218.0	11.2	7 539.0	14 210.0
1993	52 692.0	34 634.4	30 510.0	11.5	8 394.0	14 524.0
1994	76 909.0	46 759.4	33 261.0	12.4	9 281.0	14 608.0
1995	91 893.8	58 478.1	35 730.0	13.6	10 077.0	15 005.0
1996	99 595.3	67 884.6	36 454.0	14.0	10 813.0	15 733.0
1997	113 732.7	74 462.6	38 368.0	13.7	11 356.0	16 074.0
1998	119 048.0	78 345.0	38 046.0	12.5	11 670.0	16 100.0
1999	126 111.0	82 067.0	40 496.0	10.5	12 393.0	16 000.0
2000	85 673.7	89 403.5	44 452.0	10.0	13 556.0	16 300.0

8.2.3　Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有 7 个变量，分别是年份、工业总产值、国内生产总值、货物周转量、原煤、发电量和原油。我们把这些变量分别定

义为 V1、V2、V3、V4、V5、V6、V7。变量类型及长度采取系统默认方式，然后录入相关数据。相关操作我们在第 1 章中已有详细讲述。录入完成后数据如图 8.8 所示。

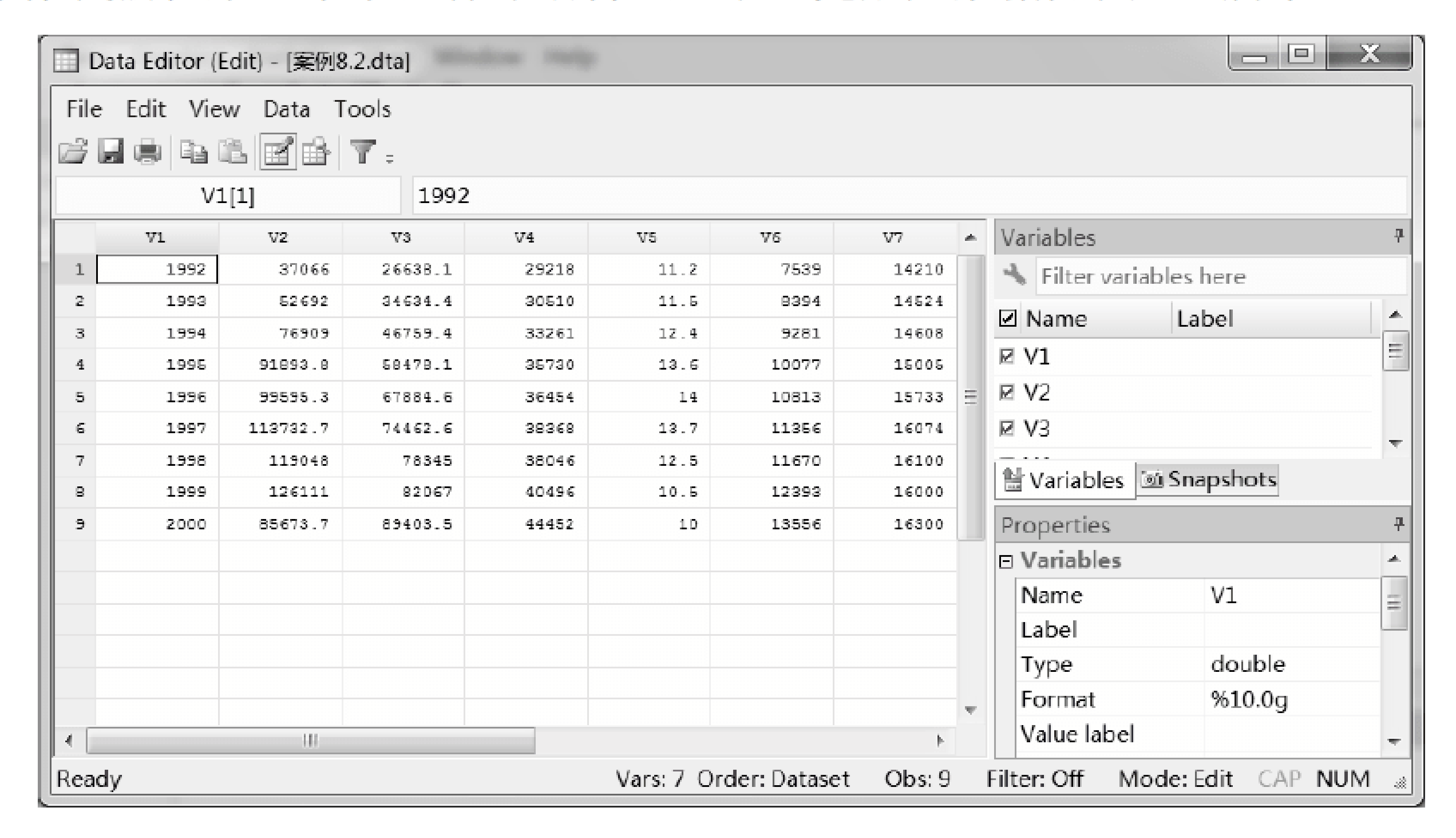


图 8.8　案例 8.2 数据

因子分析的方法有很多种，Stata 14.0 支持 4 种因子分析方法，包括主成分因子法（Principal Component Factors）、主因子法（Principal Factors）、迭代公因子方差的主因子法（Iterated Principal Factors）、最大似然因子法（Maximum Likelihood Factors）等。我们先做一下数据保存，然后开始展开分析。

1．主成分因子法

操作步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中分别输入如下命令并按键盘上的回车键进行确认。

- factor V2-V7,pcf：本命令的含义是使用主成分因子法对工业总产值、国内生产总值、货物周转量、原煤、发电量、原油变量进行因子分析。
- rotate：本命令的含义是对因子结构进行旋转。
- loadingplot,factors(2) yline(0) xline(0)：本命令的含义是绘制因子旋转后的因子载荷图。
- predict f1 f2：本命令的含义是显示因子得分系数矩阵。
- list V1 f1 f2：本命令的含义是估计因子分析后各个样本的因子得分情况。
- correlate f1 f2：本命令的含义是展示提取的主因子的相关系数矩阵。
- scoreplot,mlabel(V1) yline(0) xline(0)：本命令的含义是展示每个样本的因子得分示意图。
- estat kmo：本命令的含义是显示 KMO 检验的结果。
- screeplot：本命令的含义是绘制因子分析的碎石图。

03 设置完毕后，等待输出结果。

2．主因子法

操作步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。
02 在主界面的“Command”文本框中分别输入如下命令，并按键盘上的回车键进行确认。

- factor V2-V7,pf：本命令的含义是使用主因子法对工业总产值、国内生产总值、货物周转量、原煤、发电量、原油变量进行因子分析。
- rotate：本命令的含义是对因子结构进行旋转。
- loadingplot,factors(2) yline(0) xline(0)：本命令的含义是绘制因子旋转后的因子载荷图。
- predict f1 f2 f3 f4：本命令的含义是显示因子得分系数矩阵。
- list V1 f1 f2 f3 f4：本命令的含义是估计因子分析后各个样本的因子得分情况。
- correlate f1 f2 f3 f4：本命令的含义是展示提取的主因子的相关系数矩阵。
- scoreplot,mlabel(V1) yline(0) xline(0)：本命令的含义是展示每个样本的因子得分示意图。
- estat kmo：本命令的含义是显示 KMO 检验的结果。
- screeplot：本命令的含义是绘制因子分析的碎石图。

03 设置完毕后，等待输出结果。

3．迭代公因子方差的主因子法

操作步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。
02 在主界面的“Command”文本框中分别输入如下命令，并按键盘上的回车键进行确认。

- factor V2-V7,ipf：本命令的含义是使用迭代公因子方差的主因子法对工业总产值、国内生产总值、货物周转量、原煤、发电量、原油等变量进行因子分析。
- rotate：本命令的含义是对因子结构进行旋转。
- loadingplot,factors(2) yline(0) xline(0)：本命令的含义是绘制因子旋转后的因子载荷图。
- predict f1 f2 f3 f4 f5：本命令的含义是显示因子得分系数矩阵。
- list V1 f1 f2 f3 f4 f5：本命令的含义是估计因子分析后各个样本的因子得分情况。
- correlate f1 f2 f3 f4 f5：本命令的含义是展示提取的主因子的相关系数矩阵。
- scoreplot,mlabel(V1) yline(0) xline(0)：本命令的含义是展示每个样本的因子得分示意图。
- estat kmo：本命令的含义是显示 KMO 检验的结果。
- screeplot：本命令的含义是绘制因子分析的碎石图。

03 设置完毕后，等待输出结果。

4．最大似然因子法

操作步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。
02 在主界面的“Command”文本框中分别输入如下命令，并按键盘上的回车键进行确认。

- factor V2-V7,ml：本命令的含义是使用最大似然因子法对工业总产值、国内生产总值、

货物周转量、原煤、发电量、原油变量进行因子分析。

- rotate：本命令的含义是对因子结构进行旋转。
- loadingplot,factors(2) yline(0) xline(0)：本命令的含义是绘制因子旋转后的因子载荷图。
- predict f1 f2 f3：本命令的含义是显示因子得分系数矩阵。
- list V1 f1 f2 f3：本命令的含义是估计因子分析后各个样本的因子得分情况。
- correlate f1 f2 f3：本命令的含义是展示提取的主因子的相关系数矩阵。
- scoreplot,mlabel(V1) yline(0) xline(0)：本命令的含义是展示每个样本的因子得分示意图。
- estat kmo：本命令的含义是显示 KMO 检验的结果。
- screeplot：本命令的含义是绘制因子分析的碎石图。

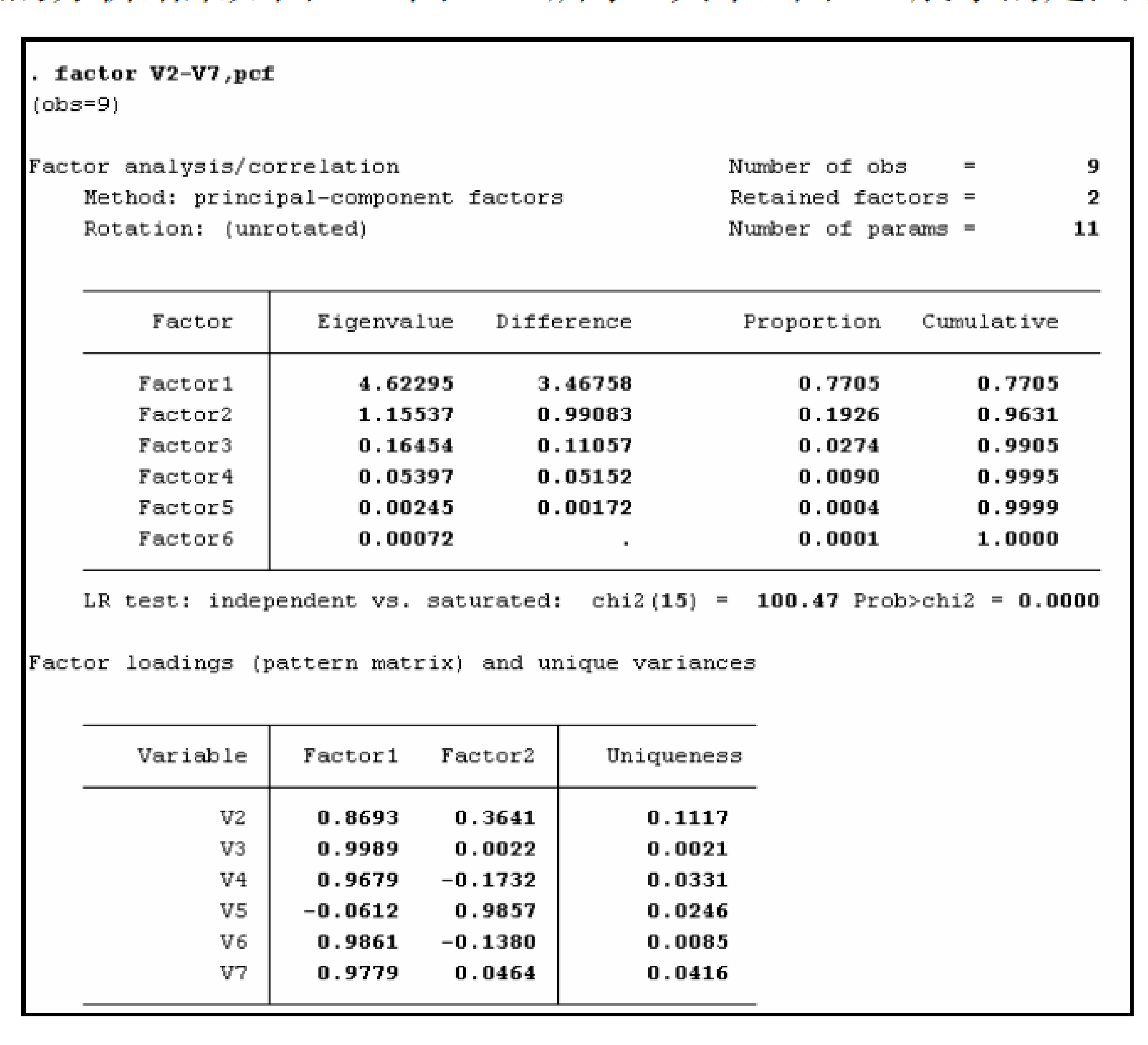
03 设置完毕后，等待输出结果。

8.2.4 结果分析

在 Stata 14.0 主界面的结果窗口可以看到如图 8.9~图 8.48 所示的分析结果。

1. 主成分因子法

主成分因子法的分析结果如图 8.9~图 8.18 所示。其中，图 8.9 展示的是因子分析的基本情况。



```
. factor V2-V7,pcf
(obs=9)

Factor analysis/correlation                   Number of obs    =        9
    Method: principal-component factors      Retained factors =        2
    Rotation: (unrotated)                     Number of params =       11
```

Factor	Eigenvalue	Difference	Proportion	Cumulative
Factor1	4.62295	3.46758	0.7705	0.7705
Factor2	1.15537	0.99083	0.1926	0.9631
Factor3	0.16454	0.11057	0.0274	0.9905
Factor4	0.05397	0.05152	0.0090	0.9995
Factor5	0.00245	0.00172	0.0004	0.9999
Factor6	0.00072	.	0.0001	1.0000

LR test: independent vs. saturated: chi2(15) = 100.47 Prob>chi2 = 0.0000

Factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Uniqueness
V2	0.8693	0.3641	0.1117
V3	0.9989	0.0022	0.0021
V4	0.9679	-0.1732	0.0331
V5	-0.0612	0.9857	0.0246
V6	0.9861	-0.1380	0.0085
V7	0.9779	0.0464	0.0416

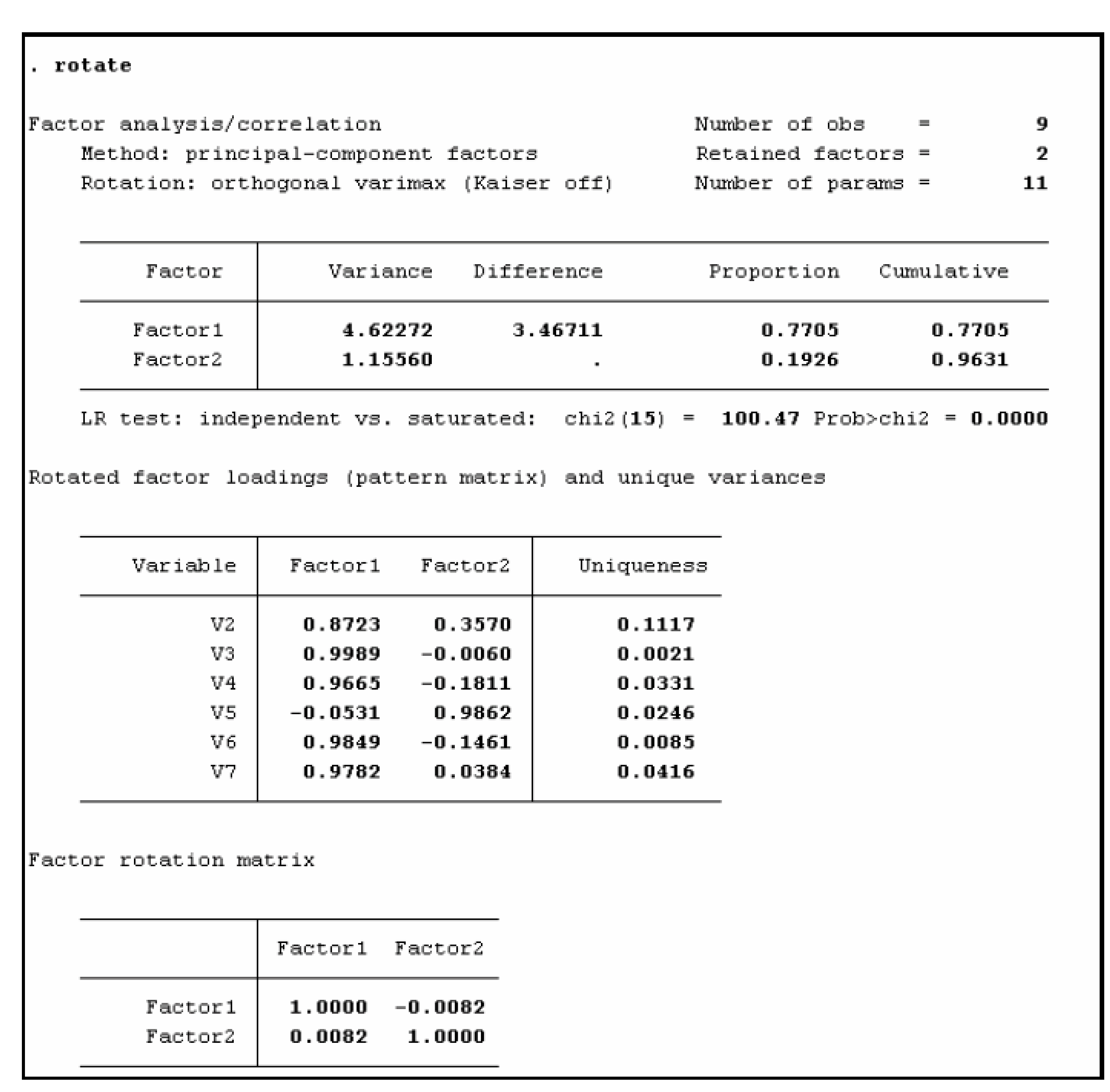
图 8.9 因子分析的基本情况

图 8.9 的上半部分说明的是因子分析模型的一般情况，从图中我们可以看出共有 9 个样本（Number of obs=9）参与了分析，提取保留的因子共有两个（Retained factors=2），模型 LR 检验的卡方值(LR test: independent vs. saturated: chi2(15))为 100.47，P 值(Prob>chi2)为 0.0000，模型非常显著。图 8.9 的上半部分最左列（Factor）说明的是因子名称，可以看出模型共提取了 6 个因子。Eigenvalue 列表示的是提取因子的特征值情况，只有前两个因子的特征值是大于

1 的，其中第 1 个因子的特征值是 4.62295，第 2 个因子的特征值是 1.15537。Proportion 列表示的是提取因子的方差贡献率，其中第 1 个因子的方差贡献率为 77.05%，第 2 个因子的方差贡献率为 19.26%。Cumulative 列表示的是提取因子的累计方差贡献率，其中前两个因子的累计方差贡献率为 96.31%。

图 8.9 的下半部分说明的是模型的因子载荷矩阵以及变量的未被解释部分。其中，Variable 列表示的是变量名称，Factor1、Factor2 两列分别说明的是提取的前两个主因子（特征值大于 1 的）对各个变量的解释程度，本例中，Factor1 主要解释的是 V2、V3、V4、V6、V7 这 5 个变量的信息，Factor2 主要解释的是 V5 变量的信息。Uniqueness 列表示变量未被提取的前两个主因子解释的部分，可以发现在舍弃其他主因子的情况下，信息的损失量是很小的。

图 8.10 展示的是对因子结构进行旋转的结果。经过学者们的研究表明，旋转操作有助于进一步简化因子结构。Stata 14.0 支持的旋转方式有两种：一种是最大方差正交旋转，一般适用于互相独立的因子或者成分，也是系统默认的情况；另外一种是 promax 斜交旋转，允许因子或者成分之间存在相关关系。此处我们选择系统默认方式，当然我们后面的操作也证明了这样做的恰当性。



```
. rotate

Factor analysis/correlation                        Number of obs    =        9
    Method: principal-component factors            Retained factors =        2
    Rotation: orthogonal varimax (Kaiser off)      Number of params =       11

    --------------------------------------------------------------------------
         Factor  |     Variance   Difference        Proportion   Cumulative
    -------------+------------------------------------------------------------
        Factor1  |      4.62272      3.46711            0.7705       0.7705
        Factor2  |      1.15560            .            0.1926       0.9631
    --------------------------------------------------------------------------
    LR test: independent vs. saturated:  chi2(15) =  100.47 Prob>chi2 = 0.0000

Rotated factor loadings (pattern matrix) and unique variances

    -------------------------------------------------
        Variable |  Factor1   Factor2 |   Uniqueness
    -------------+--------------------+--------------
              V2 |   0.8723    0.3570 |      0.1117
              V3 |   0.9989   -0.0060 |      0.0021
              V4 |   0.9665   -0.1811 |      0.0331
              V5 |  -0.0531    0.9862 |      0.0246
              V6 |   0.9849   -0.1461 |      0.0085
              V7 |   0.9782    0.0384 |      0.0416
    -------------------------------------------------

Factor rotation matrix

    --------------------------------
                 | Factor1  Factor2
    -------------+------------------
         Factor1 |  1.0000  -0.0082
         Factor2 |  0.0082   1.0000
    --------------------------------
```

图 8.10　对因子结构进行旋转

图 8.10 包括 3 部分内容，第 1 部分说明的是因子旋转模型的一般情况，从图中我们可以看出共有 9 个样本（Number of obs = 9）参与了分析，提取保留的因子共有两个（Retained factors= 2），模型 LR 检验的卡方值（LR test: independent vs. saturated: chi2(15)）为 100.47，P 值（Prob>chi2）为 0.0000，模型非常显著。最左列（Factor）说明的是因子名称，可以看出模型旋转后共提取了两个因子。Proportion 列表示的是提取因子的方差贡献率，其中第 1 个因子的方差贡献率为 77.05%，第 2 个因子的方差贡献率为 19.26%。Cumulative 列表示的是提取因子

的累计方差贡献率，其中前两个因子的累计方差贡献率为96.31%。

图8.10的第2部分说明的是模型的因子载荷矩阵以及变量的未被解释部分。其中，Variable列表示的是变量名称，Factor1、Factor2两列分别说明的是旋转提取的两个主因子对各个变量的解释程度，本例中，Factor1主要解释的是V2、V3、V4、V6、V7这5个变量的信息，Factor2主要解释的是V5变量的信息。Uniqueness列表示变量未被提取的前两个主因子解释的部分，可以发现在舍弃其他主因子的情况下，信息的损失量是很小的。

图8.10的第3部分展示的是因子旋转矩阵的一般情况，提取的两个因子不存在相关关系。

图8.11展示的是因子旋转后的因子载荷图。因子载荷图可以使用户更加直观地看出各个变量被两个因子的解释情况。

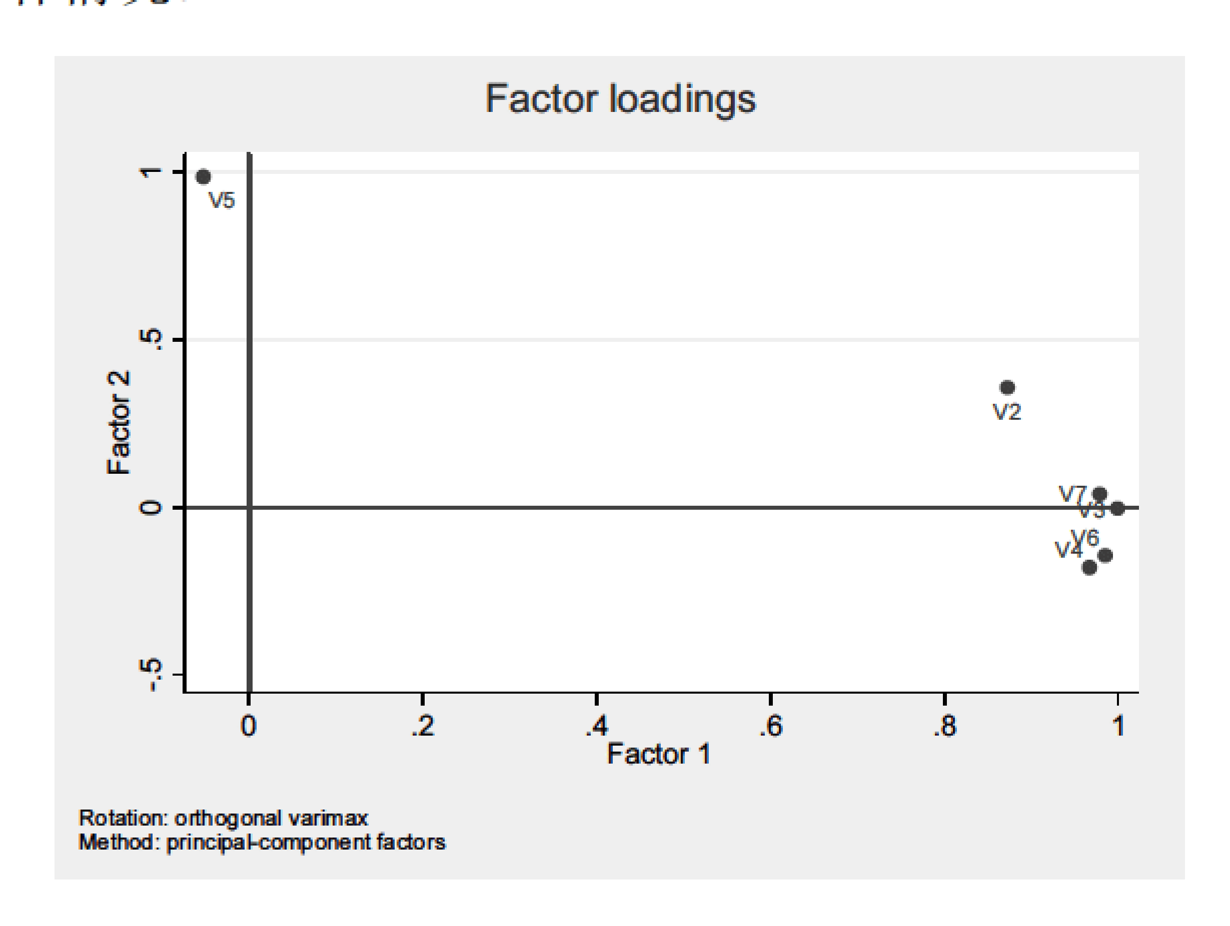


图8.11　因子载荷图

与前面的分析相同，我们发现V2、V3、V4、V6、V7这5个变量的信息主要被Factor1这一因子所解释，V5变量主要被Factor2这一因子所解释。

图8.12展示的是因子分析后各个样本的因子得分情况。因子得分的概念是通过将每个变量标准化为平均数等于0和方差等于1，然后以因子分析系数进行加权合计为每个因子构成的线性情况。以因子的方差贡献率为权数对因子进行加权求和，即可得到每个样本的因子综合得分。

根据图8.12展示的因子得分系数矩阵，我们可以写出各公因子的表达式。值得一提的是，在表达式中各个变量已经不是原始变量，而是标准化变量。

```
. predict f1 f2
(regression scoring assumed)

Scoring coefficients (method = regression; based on varimax rotated factors)
```

Variable	Factor1	Factor2
V2	0.19062	0.31358
V3	0.21609	0.00010
V4	0.20814	-0.15159
V5	-0.00625	0.85323
V6	0.21232	-0.12120
V7	0.21185	0.03840

图8.12　因子得分情况

表达式如下：

F1=0.19062*工业总产值+0.21609*国内生产总值+0.20814*货物周转量-0.00625*原煤+0.21232*发电量+0.21185*原油

F2=0.31358*工业总产值+0.0001*国内生产总值-0.15159*货物周转量+0.85323*原煤-0.1212*发电量+0.03840*原油

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 8.13 所示的因子得分数据。

Data Editor (Edit) - [案例8.2]

File Edit View Data Tools

f1[1] -1.6258838

	V1	V2	V3	V4	V5	V6	V7	f1	f2
1	1992	37066	26638.1	29218	11.2	7539	14210	-1.625884	-.7525918
2	1993	52692	34634.4	30510	11.5	8394	14524	-1.216697	-.4916763
3	1994	76909	46759.4	33261	12.4	9281	14608	-.7085732	.1536061
4	1995	91893.8	58478.1	35730	13.6	10077	15005	-.2043556	.9085287
5	1996	99595.3	67884.6	36454	14	10813	15733	.2390848	1.190549
6	1997	113732.7	74462.6	38368	13.7	11356	16074	.6272416	1.083627
7	1998	119048	78345	38046	12.5	11670	16100	.7317871	.424078
8	1999	126111	82067	40496	10.5	12393	16000	.9811613	-.8075529
9	2000	85673.7	89403.5	44452	10	13556	16300	1.176235	-1.708568

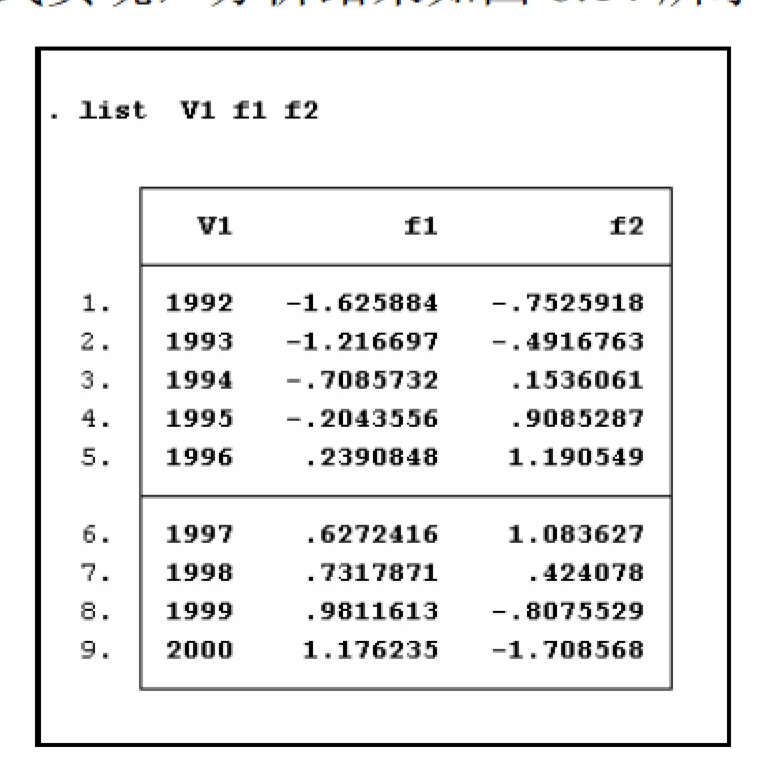
Variables: Filter variables here; Variable / Label: V1, V2, V3

Properties: Variables — Name: <2 selected>; Label: <different value; Type: float; Format: %9.0g

Ready Vars: 9 Order: Dataset Obs: 9 Filter: Off Mode: Edit CAP NUM

图 8.13 数据查看界面

当然，也可以通过命令形式实现，分析结果如图 8.14 所示。

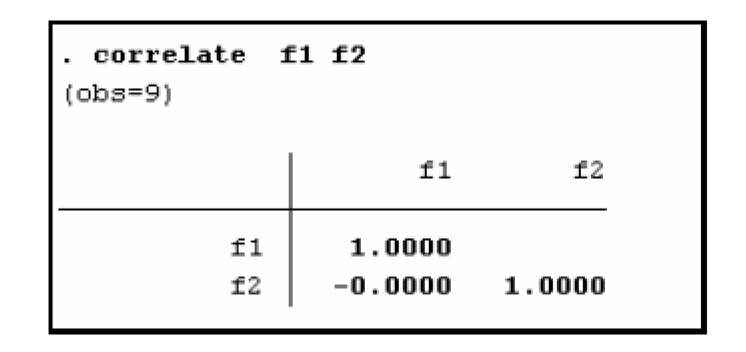


```
. list V1 f1 f2
```

	V1	f1	f2
1.	1992	-1.625884	-.7525918
2.	1993	-1.216697	-.4916763
3.	1994	-.7085732	.1536061
4.	1995	-.2043556	.9085287
5.	1996	.2390848	1.190549
6.	1997	.6272416	1.083627
7.	1998	.7317871	.424078
8.	1999	.9811613	-.8075529
9.	2000	1.176235	-1.708568

图 8.14 分析结果图

图 8.15 展示的是系统提取的两个主因子的相关系数矩阵。



```
. correlate f1 f2
(obs=9)
```

	f1	f2
f1	1.0000	
f2	-0.0000	1.0000

图 8.15 两个主因子的相关系数矩阵

从图 8.15 中可以看出，我们提取的两个主因子之间几乎没有任何相关关系，这也说明了我们在前面对因子进行旋转的操作环节中采用最大方差正交旋转方式是明智的。值得说明的是

图中 f1 与 f2 的相关系数是-0.0000，并非是不正确的，这是因为 Stata 14.0 只保留了 4 位小数所导致的，例如真实的数据有可能是-0.00001，那么结果显示的只是-0.0000。

图 8.16 展示的是每个样本的因子得分示意图。

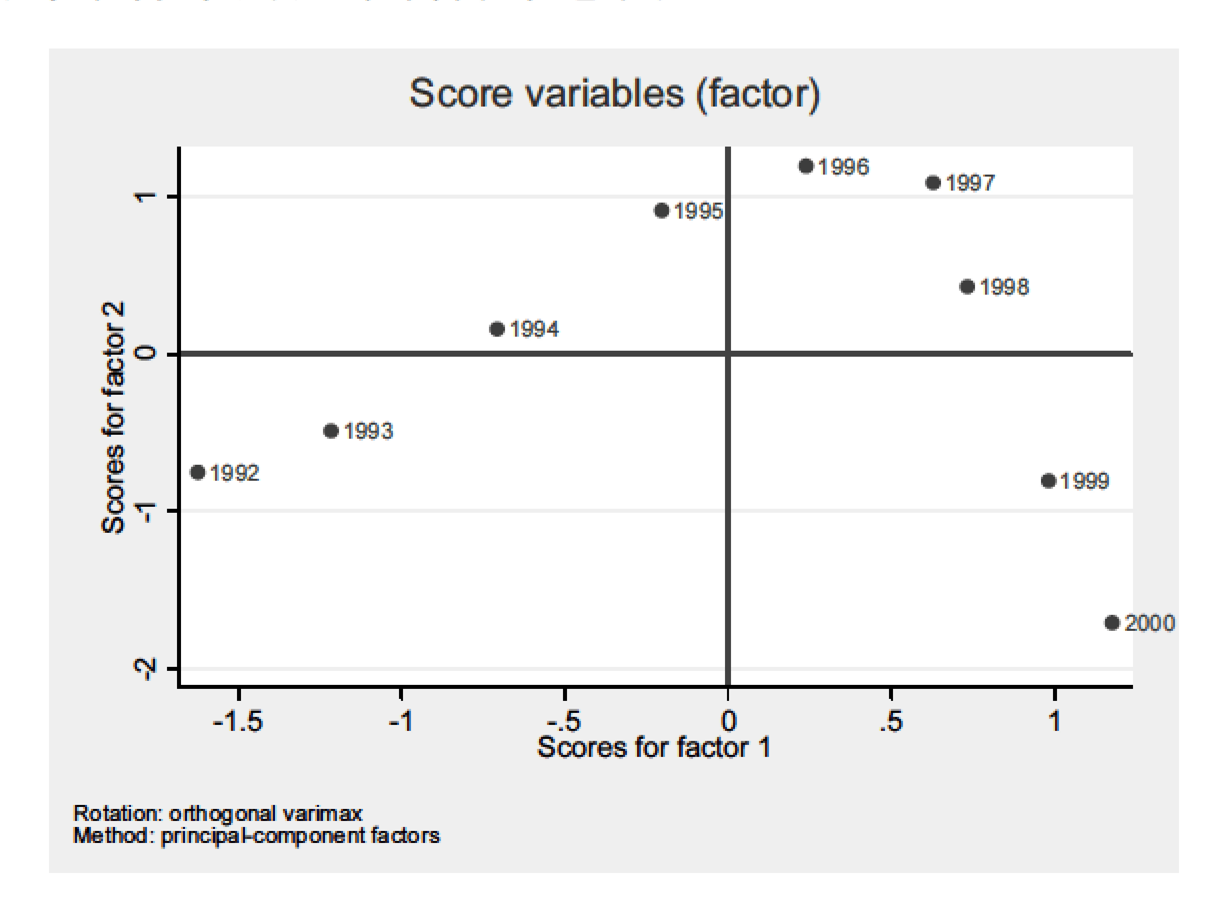


图 8.16　每个样本的因子得分示意图

从图 8.16 中可以看出，所有的样本被分到 4 个象限，其中第 1 象限包括 1996 年、1997 年、1998 年，这 3 年的两个因子得分都比较高；第 2 象限包括 1994 年、1995 年，这两年的因子 2 得分较高，而因子 1 得分较低；第 3 象限包括 1992 年、1993 年，这两年的两个因子得分都比较低；第 4 象限包括 1999 年、2000 年，这两年的因子 1 得分较高，而因子 2 得分较低。

图 8.17 展示的是本例因子分析的 KMO 检验结果。

```
. estat kmo

Kaiser-Meyer-Olkin measure of sampling adequacy
```

Variable	kmo
V2	0.6237
V3	0.6226
V4	0.7886
V5	0.1036
V6	0.6905
V7	0.7357
Overall	0.6566

图 8.17　KMO 检验结果

KMO 检验是为了判断数据是否适合进行因子分析，其取值范围是 0~1。其中，0.9~1 表示极好、0.8~0.9 表示可奖励、0.7~0.8 表示还好、0.6~0.7 表示中等、0.5~0.6 表示糟糕、0~0.5 表示不可接受。本例中总体（Overall）KMO 的取值为 0.6566，表明可以进行因子分析。各个变量的 KMO 值也大多在 0.6 以上，所以本例是比较适合因子分析的，模型的构建是有意义的。

图 8.18 展示的是本例因子分析所提取的各个因子的特征值碎石图。

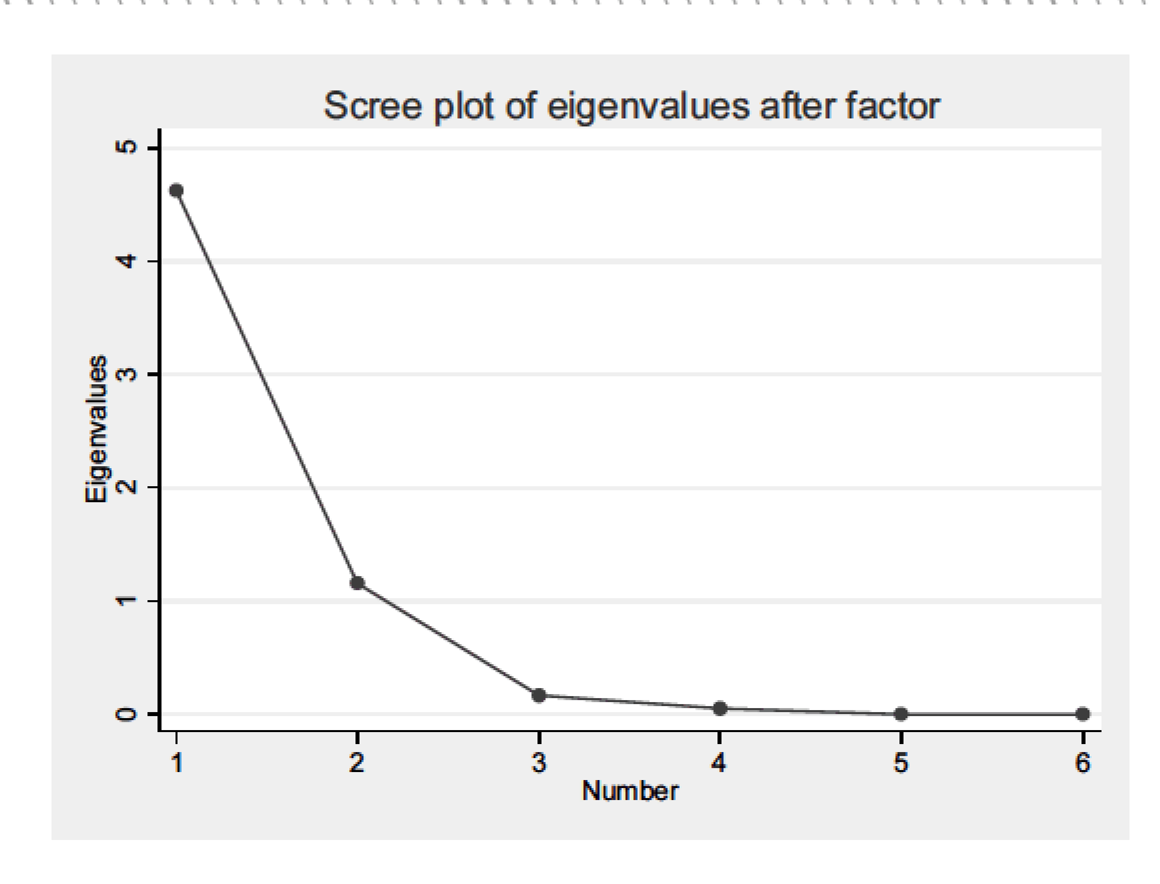
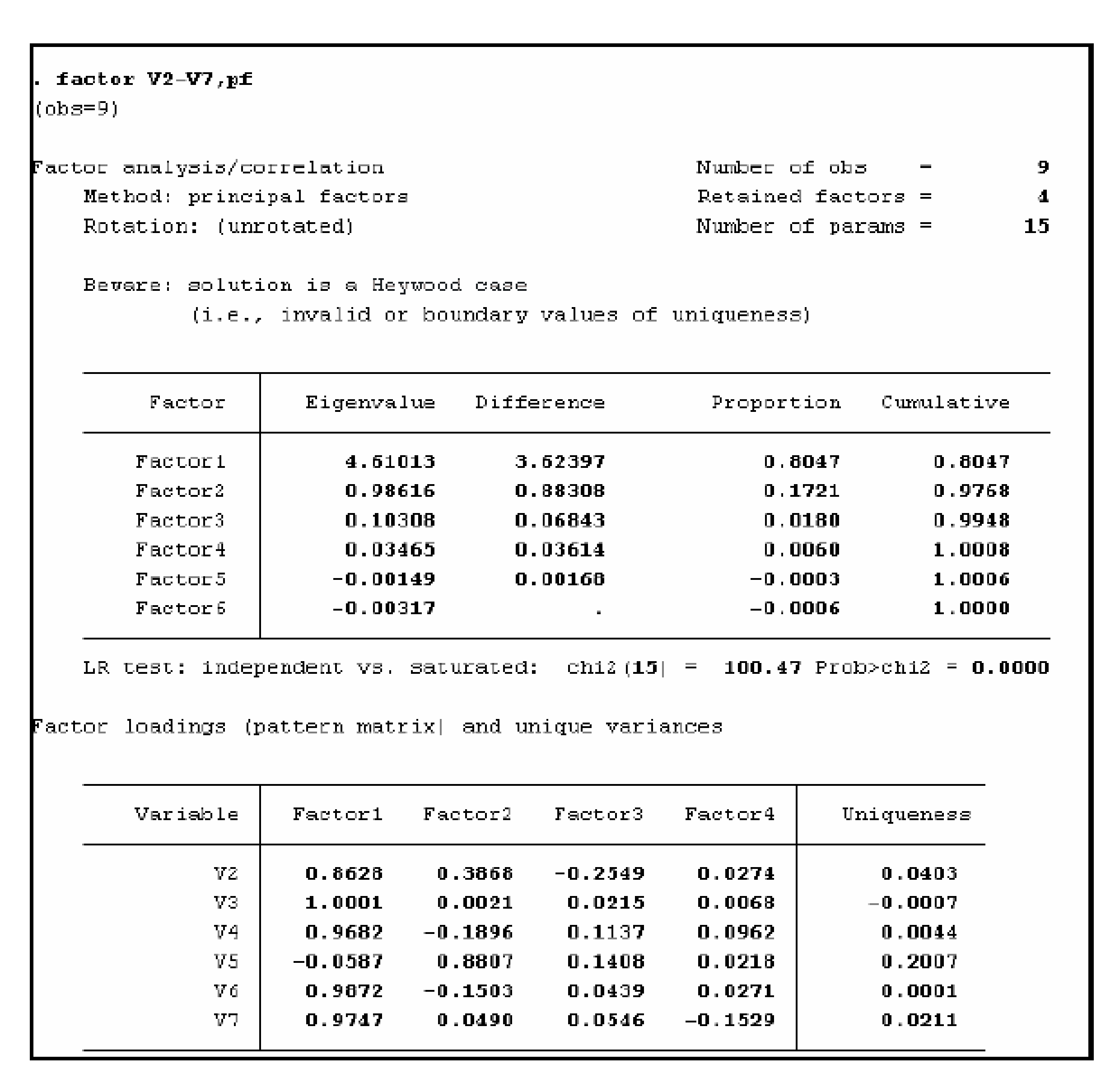


图 8.18 各个因子的特征值碎石图

通过碎石图可以非常直观地观测出提取因子的特征值的大小情况。图 8.18 的横轴表示的是系统提取因子的名称，并且已经按特征值大小进行降序排列，纵轴表示因子特征值的大小情况。从图 8.18 中可以轻松地看出本例中只有前两个因子的特征值是大于 1 的。

2．主因子法

主因子法的分析结果如图 8.19~图 8.28 所示。其中，图 8.19 展示的是因子分析的基本情况。



```
. factor V2-V7,pf
(obs=9)

Factor analysis/correlation                    Number of obs    =      9
    Method: principal factors                  Retained factors =      4
    Rotation: (unrotated)                      Number of params =     15

    Beware: solution is a Heywood case
            (i.e., invalid or boundary values of uniqueness)
```

Factor	Eigenvalue	Difference	Proportion	Cumulative
Factor1	4.61013	3.62397	0.8047	0.8047
Factor2	0.98616	0.88308	0.1721	0.9768
Factor3	0.10308	0.06843	0.0180	0.9948
Factor4	0.03465	0.03614	0.0060	1.0008
Factor5	-0.00149	0.00168	-0.0003	1.0006
Factor6	-0.00317	.	-0.0006	1.0000

```
    LR test: independent vs. saturated:  chi2(15| =  100.47 Prob>chi2 = 0.0000

Factor loadings (pattern matrix| and unique variances
```

Variable	Factor1	Factor2	Factor3	Factor4	Uniqueness
V2	0.8628	0.3868	-0.2549	0.0274	0.0403
V3	1.0001	0.0021	0.0215	0.0068	-0.0007
V4	0.9682	-0.1896	0.1137	0.0962	0.0044
V5	-0.0587	0.8807	0.1408	0.0218	0.2007
V6	0.9872	-0.1503	0.0439	0.0271	0.0001
V7	0.9747	0.0490	0.0546	-0.1529	0.0211

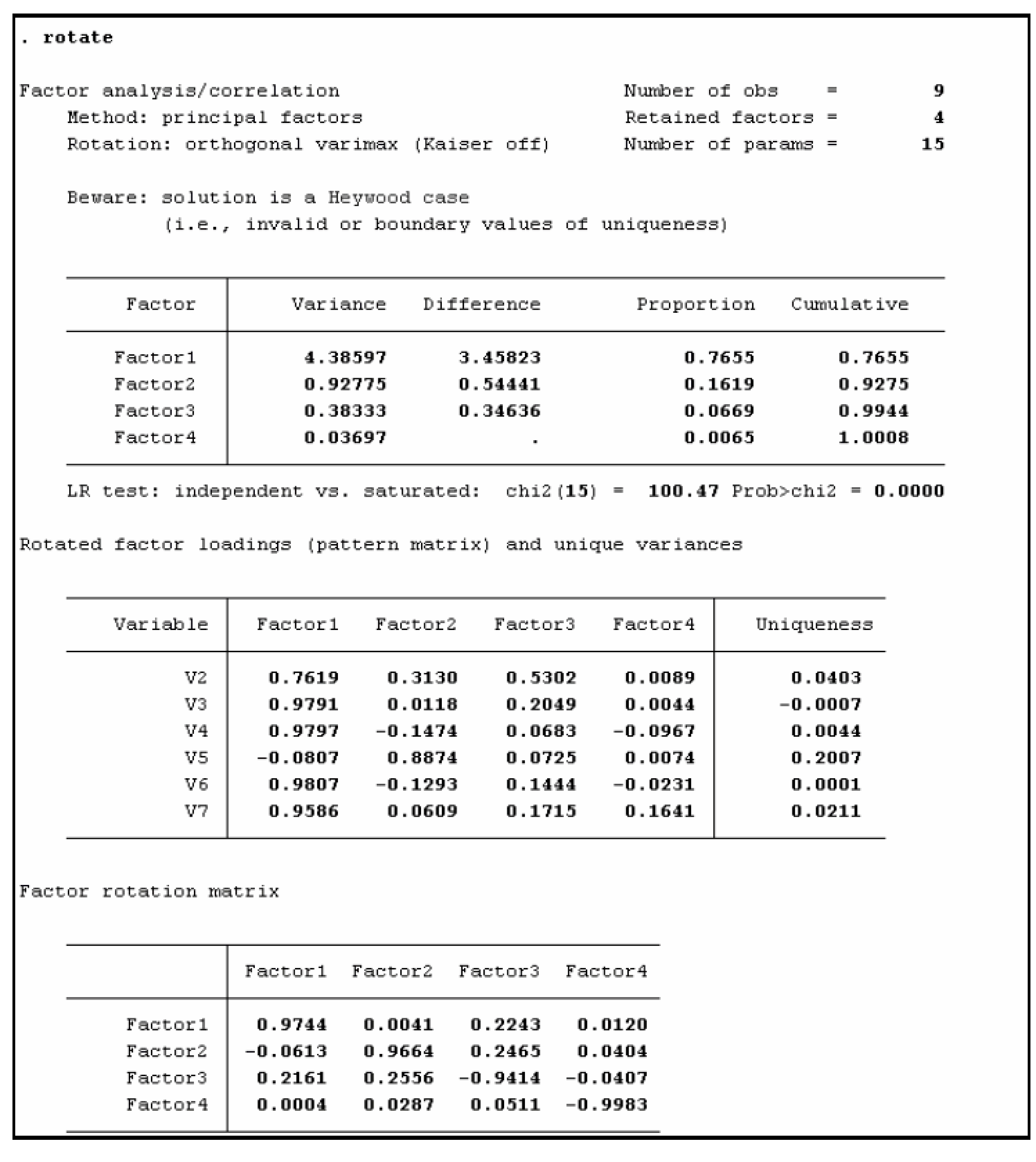
图 8.19 分析结果图

图 8.19 的上半部分说明的是因子分析模型的一般情况，从图中我们可以看出共有 9 个样本（Number of obs = 9）参与了分析，提取保留的因子共有 4 个（Retained factors = 4），模型 LR 检验的卡方值（LR test: independent vs. saturated: chi2(15)）为 100.47，P 值（Prob>chi2）为 0.0000，模型非常显著。图 8.19 的上半部分最左列（Factor）说明的是因子名称，可以看出模型共提取了 6 个因子。Eigenvalue 列表示的是提取因子的特征值情况，只有第 1 个因子的特

征值是大于 1 的，其中第 1 个因子的特征值是 4.61013，第 2 个因子的特征值是 0.98616。Proportion 列表示的是提取因子的方差贡献率，其中第 1 个因子的方差贡献率为 80.47%，第 2 个因子的方差贡献率为 17.21%。Cumulative 列表示的是提取因子的累计方差贡献率，其中前两个因子的累计方差贡献率为 97.68%。

图 8.19 的下半部分说明的是模型的因子载荷矩阵以及变量的未被解释部分。其中，Variable 列表示的是变量名称，Factor1、Factor2、Factor3、Factor4 共 4 列分别说明的是提取的 4 个主因子对各个变量的解释程度，本例中，Factor1 主要解释的是 V2、V3、V4、V6、V7 这 5 个变量的信息，Factor2 主要解释的是 V5 变量的信息。Uniqueness 列表示变量未被提取的前两个主因子解释的部分，可以发现在舍弃其他主因子的情况下，信息的损失量是很小的。

图 8.20 展示的是对因子结构进行旋转的结果。此处我们依然采用系统默认的最大方差正交旋转方式对因子结构进行旋转。



```
. rotate

Factor analysis/correlation                        Number of obs    =        9
    Method: principal factors                      Retained factors =        4
    Rotation: orthogonal varimax (Kaiser off)      Number of params =       15

    Beware: solution is a Heywood case
            (i.e., invalid or boundary values of uniqueness)

    --------------------------------------------------------------------------
         Factor  |     Variance   Difference        Proportion   Cumulative
    -------------+------------------------------------------------------------
        Factor1  |      4.38597      3.45823            0.7655       0.7655
        Factor2  |      0.92775      0.54441            0.1619       0.9275
        Factor3  |      0.38333      0.34636            0.0669       0.9944
        Factor4  |      0.03697            .            0.0065       1.0008
    --------------------------------------------------------------------------
    LR test: independent vs. saturated:  chi2(15) =  100.47 Prob>chi2 = 0.0000

Rotated factor loadings (pattern matrix) and unique variances

    -------------------------------------------------------------------
        Variable |  Factor1   Factor2   Factor3   Factor4 |   Uniqueness
    -------------+----------------------------------------+-----------
              V2 |   0.7619    0.3130    0.5302    0.0089 |     0.0403
              V3 |   0.9791    0.0118    0.2049    0.0044 |    -0.0007
              V4 |   0.9797   -0.1474    0.0683   -0.0967 |     0.0044
              V5 |  -0.0807    0.8874    0.0725    0.0074 |     0.2007
              V6 |   0.9807   -0.1293    0.1444   -0.0231 |     0.0001
              V7 |   0.9586    0.0609    0.1715    0.1641 |     0.0211
    -------------------------------------------------------------------

Factor rotation matrix

    --------------------------------------------------
                 | Factor1  Factor2  Factor3  Factor4
    -------------+------------------------------------
         Factor1 |  0.9744   0.0041   0.2243   0.0120
         Factor2 | -0.0613   0.9664   0.2465   0.0404
         Factor3 |  0.2161   0.2556  -0.9414  -0.0407
         Factor4 |  0.0004   0.0287   0.0511  -0.9983
    --------------------------------------------------
```

图 8.20　分析结果图

图 8.20 包括 3 部分内容，第 1 部分说明的是因子旋转模型的一般情况，从图中我们可以看出共有 9 个样本（Number of obs = 9）参与了分析，提取保留的因子共有 4 个（Retained factors = 4），模型 LR 检验的卡方值（LR test: independent vs. saturated: chi2(15)）为 100.47，P 值（Prob>chi2）为 0.0000，模型非常显著。最左列（Factor）说明的是因子名称，可以看出模型旋转后共提取了 4 个因子。Proportion 列表示的是提取因子的方差贡献率，其中第 1 个因子的方差贡献率为 76.55%，第 2 个因子的方差贡献率为 16.19%。Cumulative 列表示的是提取因子

的累计方差贡献率，其中前两个因子的累计方差贡献率为 92.75%。

图 8.20 的第 2 部分说明的是模型的因子载荷矩阵以及变量的未被解释部分。其中，Variable 列表示的是变量名称，Factor1、Factor2 两列分别说明的是旋转提取的两个主因子对各个变量的解释程度，本例中，Factor1 主要解释的是 V2、V3、V4、V6、V7 这 5 个变量的信息，Factor2 主要解释的是 V5 变量的信息。Uniqueness 列表示变量未被提取的前两个主因子解释的部分，可以发现在舍弃其他主因子的情况下，信息的损失量是很小的。

图 8.20 的第 3 部分展示的是因子旋转矩阵的一般情况，提取的 4 个因子相关关系很弱。

图 8.21 展示的是因子旋转后的因子载荷图。此处我们通过 Factor 选项控制了因子的数目，本因子载荷图可以使用户更加直观地看出各个变量被前两个因子解释的情况。

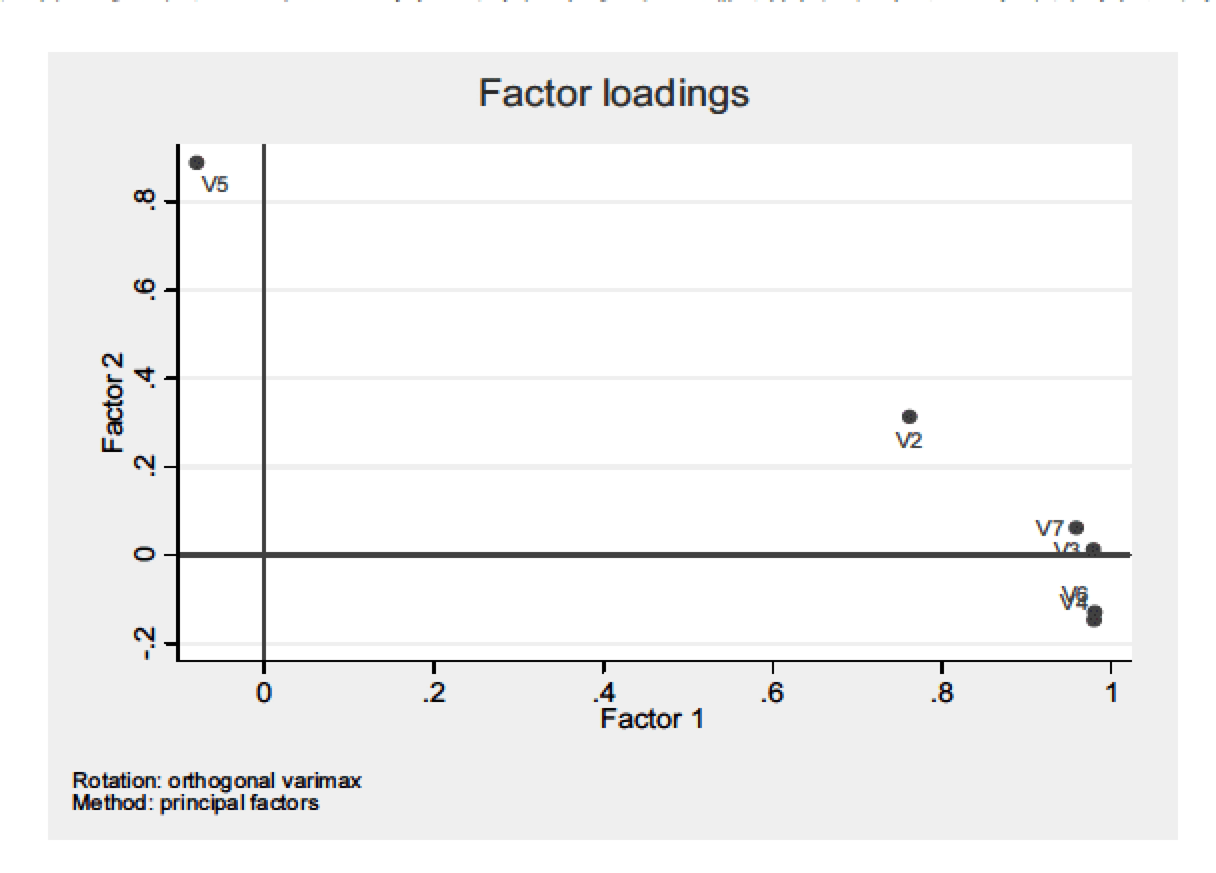


图 8.21　旋转后的因子载荷图

与前面的分析相同，我们发现 V2、V3、V4、V6、V7 这 5 个变量的信息主要被 Factor1 这一因子所解释，V5 变量主要被 Factor2 这一因子所解释。

图 8.22 展示的是因子分析后各个样本的因子得分情况。

```
. predict f1 f2 f3 f4
(regression scoring assumed)

Scoring coefficients (method = regression; based on varimax rotated factors)
```

Variable	Factor1	Factor2	Factor3	Factor4
V2	-0.36964	-0.86144	1.16737	-1.08005
V3	1.54910	10.02019	0.70491	1.65788
V4	0.55537	1.87638	-4.08809	-5.09758
V5	0.04255	0.03824	-0.31572	-0.10529
V6	-0.92229	-1.1e+01	3.95317	2.58399
V7	0.12414	-0.40643	-1.48796	1.70736

图 8.22　各个样本的因子得分情况

根据图 8.22 展示的因子得分系数矩阵，我们可以写出各公因子的表达式。值得一提的是，在表达式中各个变量已经不是原始变量，而是标准化变量。

表达式如下：

F1=-0.36964*工业总产值+1.54910*国内生产总值+0.55537*货物周转量+0.04255*原煤-0.92229*发电量+0.12414*原油

F2=-0.86144*工业总产值+10.02019*国内生产总值+1.87638*货物周转量+0.03824*原煤

-1.1e+01*发电量-0.40643*原油

F3=1.16737*工业总产值+0.70491*国内生产总值-4.08809*货物周转量-0.31572*原煤 +3.95317*发电量-1.48796*原油

F4=-1.08005 *工业总产值+1.65788*国内生产总值-5.09758*货物周转量-0.18529*原煤 +2.58399*发电量+1.70736*原油

我们选择“Data” | “Data Editor” | “Data Editor(Browse)” 命令，进入数据查看界面，可以看到如图 8.23 所示的因子得分数据。

Data Editor (Edit) - [案例8.2]

File Edit View Data Tools

f1[1] -1.4430785

	V1	V2	V3	V4	V5	V6	V7	f1	f2	f3	f4
1	1992	37066	26638.1	29218	11.2	7539	14210	-1.443079	.0581607	-.9417189	.2679824
2	1993	52692	34634.4	30510	11.5	8394	14524	-1.272485	-1.169193	-.0742355	.7067811
3	1994	76909	46759.4	33261	12.4	9281	14608	-.7808981	-.2360872	.3783891	-.9217796
4	1995	91893.8	58478.1	35730	13.6	10077	15005	-.1362868	1.025807	-.1311406	-1.445611
5	1996	99595.3	67884.6	36454	14	10813	15733	.2893678	.9075351	-.0733339	.6936998
6	1997	113732.7	74462.6	38368	13.7	11356	16074	.585381	1.034194	-.395167	.1363798
7	1998	119048	78345	38046	12.5	11670	16100	.5756679	.7267231	1.065623	1.207648
8	1999	126111	82067	40496	10.5	12393	16000	.6156501	-.866867	1.473271	-.3559471
9	2000	85673.7	89403.5	44452	10	13556	16300	1.566681	-1.480272	-1.301687	-.2891538

Variables: Filter variables here; Variable, Label; V1; V2; V3

Properties: Variables; Name <4 selected>; Label <different value; Type float; Format %9.0g

Ready Vars: 11 Order: Dataset Obs: 9 Filter: Off Mode: Edit CAP NUM

图 8.23 数据查看界面

这一点也可以通过命令形式实现，如图 8.24 所示。

```
. list V1 f1 f2 f3 f4

     +----------------------------------------------------------+
     |   V1          f1          f2          f3          f4     |
     |----------------------------------------------------------|
  1. | 1992   -1.443079    .0581607   -.9417189    .2679824     |
  2. | 1993   -1.272485   -1.169193   -.0742355    .7067811     |
  3. | 1994   -.7808981   -.2360872    .3783891   -.9217796     |
  4. | 1995   -.1362868    1.025807   -.1311406   -1.445611     |
  5. | 1996    .2893678    .9075351   -.0733339    .6936998     |
     |----------------------------------------------------------|
  6. | 1997     .585381    1.034194    -.395167    .1363798     |
  7. | 1998    .5756679    .7267231    1.065623    1.207648     |
  8. | 1999    .6156501    -.866867    1.473271   -.3559471     |
  9. | 2000    1.566681   -1.480272   -1.301687   -.2891538     |
     +----------------------------------------------------------+
```

图 8.24 通过命令形式实现

图 8.25 展示的是系统提取的 4 个主因子的相关系数矩阵。

```
. correlate f1 f2 f3 f4
(obs=9)

             |       f1       f2       f3       f4
-------------+------------------------------------
          f1 |   1.0000
          f2 |  -0.0150   1.0000
          f3 |   0.0576   0.1084   1.0000
          f4 |  -0.0098   0.0463   0.0960   1.0000
```

图 8.25 4 个主因子的相关系数矩阵

从图 8.25 中可以看出，我们提取的 4 个主因子之间几乎没有什么相关关系，这也说明了我们在前面对因子进行旋转的操作环节中采用最大方差正交旋转方式是明智的。

图 8.26 展示的是每个样本的因子得分示意图。

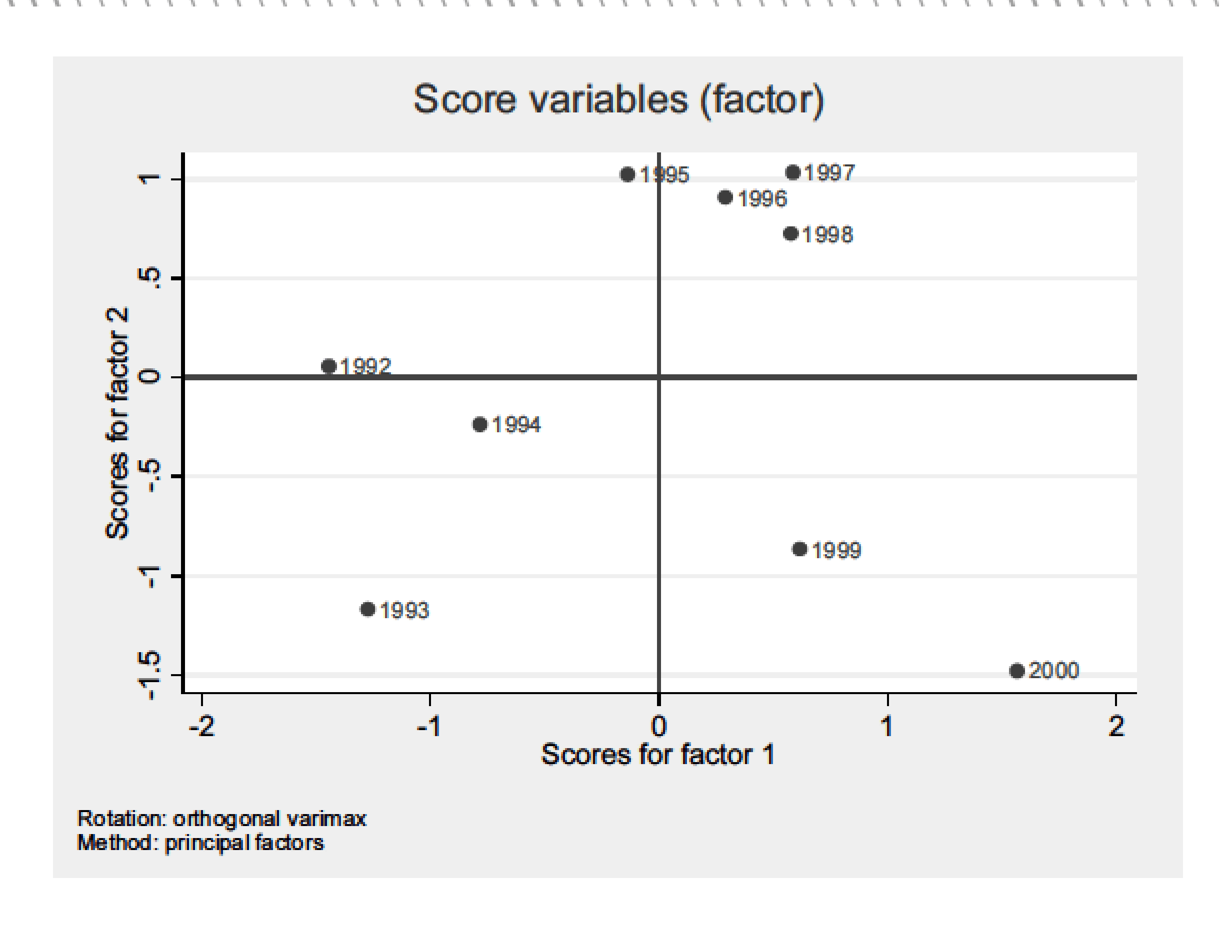


图 8.26　因子得分示意图

从图 8.26 中可以看出，所有的样本被分到 4 个象限，其中第 1 象限包括 1996 年、1997 年、1998 年，这 3 年的两个因子得分都比较高；第 2 象限包括 1992 年、1995 年，这两年的因子 2 得分较高，而因子 1 得分较低；第 3 象限包括 1993 年、1994 年，这两年的两个因子得分都比较低；第 4 象限包括 1999 年、2000 年，这两年的因子 1 得分较高，而因子 2 得分较低。

图 8.27 展示的是本例因子分析的 KMO 检验结果。

```
. estat kmo

Kaiser-Meyer-Olkin measure of sampling adequacy

    ---------------------------
        Variable |     kmo
    -------------+-------------
              V2 |   0.6237
              V3 |   0.6226
              V4 |   0.7886
              V5 |   0.1036
              V6 |   0.6905
              V7 |   0.7357
    -------------+-------------
         Overall |   0.6566
    ---------------------------
```

图 8.27　KMO 检验结果

KMO 检验的结果与前面是一致的。

图 8.28 展示的是本例因子分析所提取的各个因子的特征值碎石图。

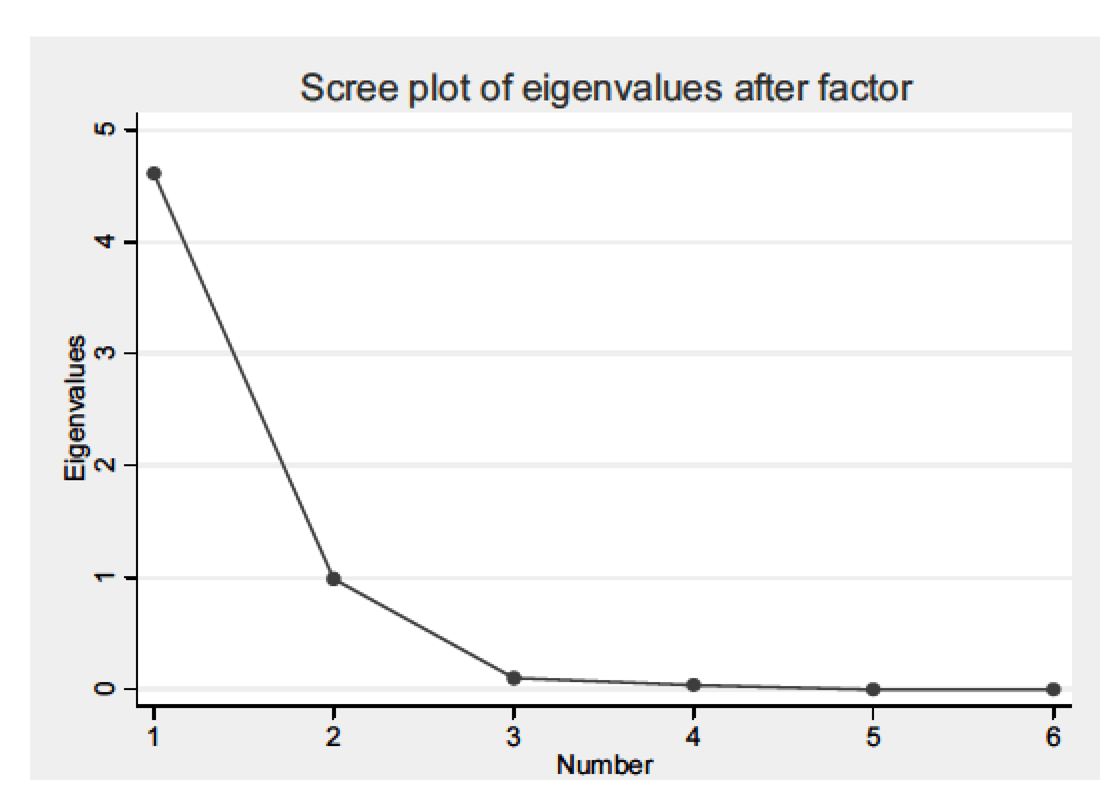
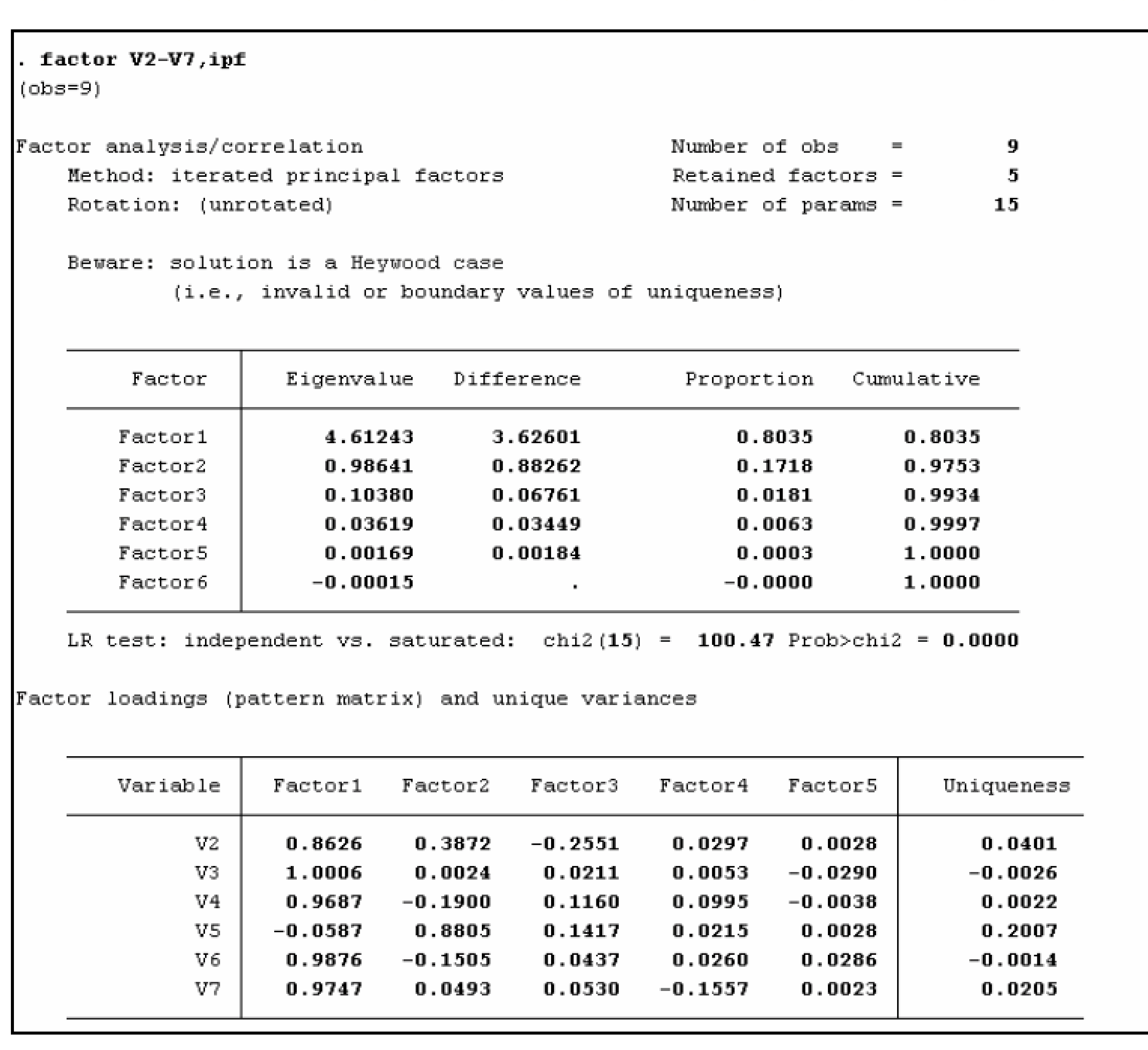


图 8.28　各个因子的特征值碎石图

从图 8.28 中可以轻松地看出本例中只有第 1 个因子的特征值是明显大于 1 的，第 2 个因子的特征值是接近于 1 的。

3．迭代公因子方差的主因子法

分析结果如图 8.29~图 8.38 所示。其中，图 8.29 展示的是因子分析的基本情况。



```
. factor V2-V7,ipf
(obs=9)

Factor analysis/correlation                    Number of obs    =          9
    Method: iterated principal factors         Retained factors =          5
    Rotation: (unrotated)                      Number of params =         15

    Beware: solution is a Heywood case
            (i.e., invalid or boundary values of uniqueness)

    --------------------------------------------------------------------------
         Factor  |   Eigenvalue   Difference        Proportion   Cumulative
    -------------+------------------------------------------------------------
        Factor1  |      4.61243      3.62601            0.8035       0.8035
        Factor2  |      0.98641      0.88262            0.1718       0.9753
        Factor3  |      0.10380      0.06761            0.0181       0.9934
        Factor4  |      0.03619      0.03449            0.0063       0.9997
        Factor5  |      0.00169      0.00184            0.0003       1.0000
        Factor6  |     -0.00015            .           -0.0000       1.0000
    --------------------------------------------------------------------------
    LR test: independent vs. saturated:  chi2(15) =  100.47 Prob>chi2 = 0.0000

Factor loadings (pattern matrix) and unique variances

    -------------------------------------------------------------------------------
        Variable |  Factor1   Factor2   Factor3   Factor4   Factor5 |   Uniqueness
    -------------+--------------------------------------------------+--------------
              V2 |   0.8626    0.3872   -0.2551    0.0297    0.0028 |      0.0401
              V3 |   1.0006    0.0024    0.0211    0.0053   -0.0290 |     -0.0026
              V4 |   0.9687   -0.1900    0.1160    0.0995   -0.0038 |      0.0022
              V5 |  -0.0587    0.8805    0.1417    0.0215    0.0028 |      0.2007
              V6 |   0.9876   -0.1505    0.0437    0.0260    0.0286 |     -0.0014
              V7 |   0.9747    0.0493    0.0530   -0.1557    0.0023 |      0.0205
    -------------------------------------------------------------------------------
```

图 8.29　因子分析的基本情况

图 8.29 的上半部分说明的是因子分析模型的一般情况，从图中我们可以看出共有 9 个样本（Number of obs = 9）参与了分析，提取保留的因子共有 5 个（Retained factors = 5），模型 LR 检验的卡方值（LR test: independent vs. saturated: chi2(15)）为 100.47，P 值（Prob>chi2）为 0.0000，模型非常显著。图 8.29 的上半部分最左列（Factor）说明的是因子名称，可以看出模型共提取了 6 个因子。Eigenvalue 列表示的是提取因子的特征值情况，只有第 1 个因子的特征值是大于 1 的，其中第 1 个因子的特征值是 4.61243，第 2 个因子的特征值是 0.98641。Proportion 列表示的是提取因子的方差贡献率，其中第 1 个因子的方差贡献率为 80.35%，第 2 个因子的方差贡献率为 17.18%。Cumulative 列表示的是提取因子的累计方差贡献率，其中前两个因子的累计方差贡献率为 97.53%。

图 8.29 的下半部分说明的是模型的因子载荷矩阵以及变量的未被解释部分。其中，Variable 列表示的是变量名称，Factor1、Factor2、Factor3、Factor4、Factor5 这 5 列分别说明的是提取的 5 个主因子对各个变量的解释程度，本例中，Factor1 主要解释的是 V2、V3、V4、V6、V7 这 5 个变量的信息，Factor2 主要解释的是 V5 变量的信息。Uniqueness 列表示变量未被提取的前两个主因子解释的部分，可以发现在舍弃其他主因子的情况下，信息的损失量是很小的。

图 8.30 展示的是对因子结构进行旋转的结果。此处我们依然采用系统默认的最大方差正交旋转方式对因子结构进行旋转。

```
. rotate

Factor analysis/correlation                        Number of obs    =        9
    Method: iterated principal factors             Retained factors =        5
    Rotation: orthogonal varimax (Kaiser off)      Number of params =       15

    Beware: solution is a Heywood case
            (i.e., invalid or boundary values of uniqueness)

    --------------------------------------------------------------------------
         Factor  |     Variance   Difference        Proportion   Cumulative
    -------------+------------------------------------------------------------
        Factor1  |      4.38428      3.45713            0.7638       0.7638
        Factor2  |      0.92715      0.53849            0.1615       0.9253
        Factor3  |      0.38866      0.34998            0.0677       0.9930
        Factor4  |      0.03868      0.03694            0.0067       0.9997
        Factor5  |      0.00174            .            0.0003       1.0000
    --------------------------------------------------------------------------
    LR test: independent vs. saturated:  chi2(15) =  100.47 Prob>chi2 = 0.0000

Rotated factor loadings (pattern matrix) and unique variances

    ---------------------------------------------------------------------------------
        Variable |  Factor1   Factor2   Factor3   Factor4   Factor5 |   Uniqueness
    -------------+--------------------------------------------------+----------------
              V2 |   0.7604    0.3126    0.5328    0.0094   -0.0004 |      0.0401
              V3 |   0.9791    0.0119    0.2073    0.0059    0.0289 |     -0.0026
              V4 |   0.9806   -0.1468    0.0686   -0.1004    0.0000 |      0.0022
              V5 |  -0.0812    0.8873    0.0732    0.0076    0.0011 |      0.2007
              V6 |   0.9807   -0.1292    0.1466   -0.0210   -0.0301 |     -0.0014
              V7 |   0.9579    0.0610    0.1739    0.1673    0.0015 |      0.0205
    ---------------------------------------------------------------------------------

Factor rotation matrix

    --------------------------------------------------------------
                 |  Factor1  Factor2  Factor3  Factor4  Factor5
    -------------+------------------------------------------------
         Factor1 |   0.9740   0.0042   0.2262   0.0127   0.0000
         Factor2 |  -0.0623   0.9658   0.2481   0.0413   0.0055
         Factor3 |   0.2179   0.2576  -0.9402  -0.0471  -0.0036
         Factor4 |  -0.0005   0.0280   0.0576  -0.9977  -0.0234
         Factor5 |  -0.0011   0.0038   0.0034   0.0237  -0.9997
    --------------------------------------------------------------
```

图 8.30 对因子结构进行旋转

图 8.30 包括 3 部分内容，第 1 部分说明的是因子旋转模型的一般情况，从图中我们可以看出共有 9 个样本（Number of obs = 9）参与了分析，提取保留的因子共有 5 个（Retained factors = 5），模型 LR 检验的卡方值（LR test: independent vs. saturated: chi2(15)）为 100.47，P 值（Prob>chi2）为 0.0000，模型非常显著。最左列（Factor）说明的是因子名称，可以看出模型旋转后共提取了 5 个因子。Proportion 列表示的是提取因子的方差贡献率，其中第 1 个因子的方差贡献率为 76.38%，第 2 个因子的方差贡献率为 16.15%。Cumulative 列表示的是提取因子的累计方差贡献率，其中前两个因子的累计方差贡献率为 92.53%。

图 8.30 的第 2 部分说明的是模型的因子载荷矩阵以及变量的未被解释部分。其中，Variable 列表示的是变量名称，Factor1、Factor2 两列分别说明的是旋转提取的两个主因子对各个变量的解释程度，本例中，Factor1 主要解释的是 V2、V3、V4、V6、V7 这 5 个变量的信息，Factor2 主要解释的是 V5 变量的信息。Uniqueness 列表示变量未被提取的前两个主因子解释的部分，可以发现在舍弃其他主因子的情况下，信息的损失量是很小的。

图 8.30 的第 3 部分展示的是因子旋转矩阵的一般情况，提取的 5 个因子相关关系很弱。

图 8.31 展示的是因子旋转后的因子载荷图。此处我们通过 Factor 选项控制了因子的数目，本因子载荷图可以使用户更加直观地看出各个变量被前两个因子解释的情况。

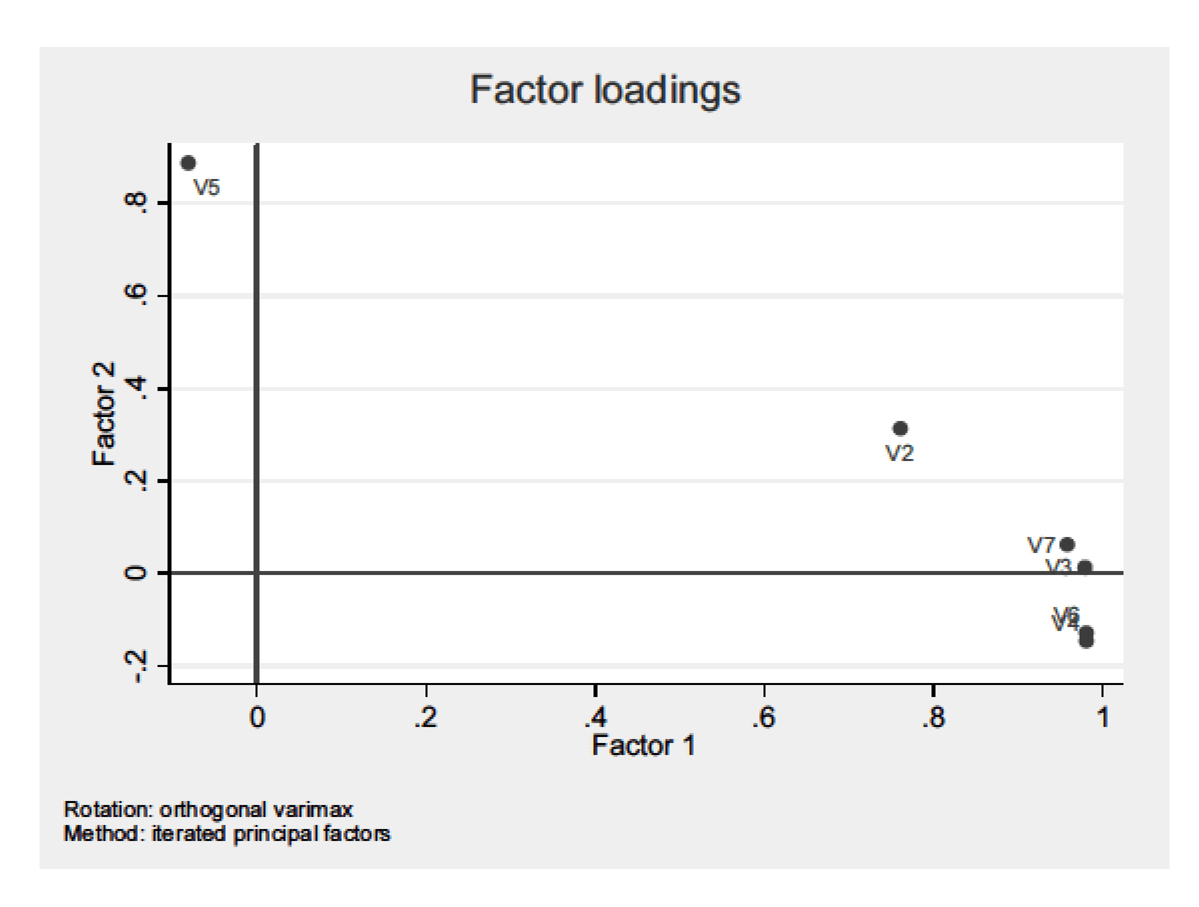


图 8.31　旋转后的因子载荷图

与前面的分析相同，我们发现 V2、V3、V4、V6、V7 这 5 个变量的信息主要被 Factor1 这一因子所解释，V5 变量主要被 Factor2 这一因子所解释。

图 8.32 展示的是因子分析后各个样本的因子得分情况。

```
. predict f1 f2 f3 f4 f5
(regression scoring assumed)

Scoring coefficients (method = regression; based on varimax rotated factors)
```

Variable	Factor1	Factor2	Factor3	Factor4	Factor5
V2	-0.39572	-0.85513	1.10138	-1.19566	-5.07934
V3	1.77032	10.03571	0.93258	1.58090	42.31499
V4	0.61805	1.96072	-4.40549	-6.06184	2.58152
V5	0.02908	0.03244	-0.30787	-0.11809	-2.66822
V6	-1.16279	-1.1e+01	4.19422	3.90009	-3.7e+01
V7	0.10074	-0.38844	-1.58820	1.52105	-4.53620

图 8.32　各个样本的因子得分情况

根据图 8.32 展示的因子得分系数矩阵，我们可以写出各公因子的表达式。值得一提的是，在表达式中各个变量已经不是原始变量，而是标准化变量。

表达式如下：

F1=-0.39572*工业总产值+1.77032*国内生产总值+ 0.61805*货物周转量+0.02908*原煤 -1.16279*发电量+0.10074*原油

F2=-0.85513*工业总产值+10.03571*国内生产总值+1.96072 *货物周转量+0.03244*原煤 -1.1e+01*发电量-0.38844*原油

F3=1.10138*工业总产值+0.93258*国内生产总值-4.40549*货物周转量-0.30787*原煤 +4.19422*发电量-1.58820*原油

F4=-1.19566 *工业总产值+1.58090*国内生产总值-6.06184*货物周转量-0.11809*原煤 +3.90009*发电量+1.52105*原油

F5=-5.07934 *工业总产值+42.31499*国内生产总值+2.58152*货物周转量-2.66822*原煤 -3.7e+01*发电量-4.53620*原油

我们选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 8.33 所示的因子得分数据。

Data Editor (Edit) – [案例8.2]

File Edit View Data Tools

f1[1] -1.4276912

	V1	V2	V3	V4	V5	V6	V7	f1	f2	f3	f4	f5
1	1992	37066	26638.1	29218	11.2	7539	14210	-1.427691	.0664599	-.9624259	.1830451	2.36591
2	1993	52692	34634.4	30510	11.5	8394	14524	-1.29143	-1.177428	-.0627082	.7970472	-2.66172
3	1994	76909	46759.4	33261	12.4	9281	14608	-.7838829	-.2400735	.3855194	-.8939711	-.817812
4	1995	91893.8	58478.1	35730	13.6	10077	15005	-.123623	1.030422	-.1423163	-1.50698	1.06167
5	1996	99595.3	67884.6	36454	14	10813	15733	.2832312	.9013298	-.0484685	.7751203	-.474841
6	1997	113732.7	74462.6	38368	13.7	11356	16074	.5836045	1.043757	-.4360728	.0325833	-.797110
7	1998	119048	78345	38046	12.5	11670	16100	.57557	.7200586	1.10393	1.286241	1.73996
8	1999	126111	82067	40496	10.5	12393	16000	.610529	-.8662681	1.464436	-.3873031	-.389408
9	2000	85673.7	89403.5	44452	10	13556	16300	1.573692	-1.478257	-1.301894	-.2857831	-.026660

Ready　Vars: 12　Order: Dataset　Obs: 9　Filter: Off　Mode: Edit

图 8.33　数据查看界面

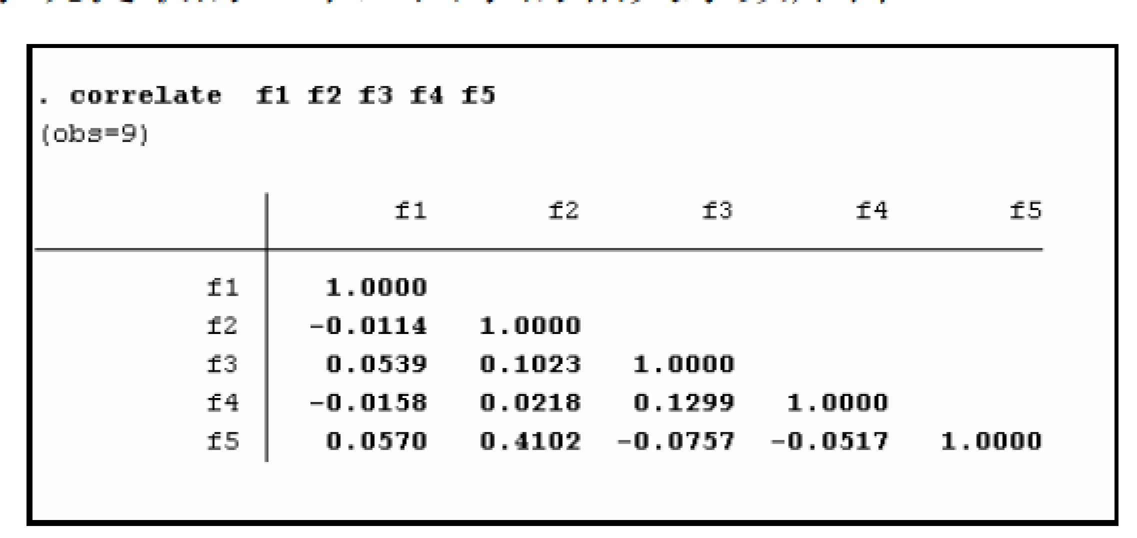
这一点也可以通过命令形式实现，如图 8.34 所示。

```
. list V1 f1 f2 f3 f4 f5
```

	V1	f1	f2	f3	f4	f5
1.	1992	-1.427691	.0664599	-.9624259	.1830451	2.365919
2.	1993	-1.29143	-1.177428	-.0627082	.7970472	-2.661722
3.	1994	-.7838829	-.2400735	.3855194	-.8939711	-.8178126
4.	1995	-.123623	1.030422	-.1423163	-1.50698	1.061674
5.	1996	.2832312	.9013298	-.0484685	.7751203	-.4748417
6.	1997	.5836045	1.043757	-.4360728	.0325833	-.7971101
7.	1998	.57557	.7200586	1.10393	1.286241	1.739962
8.	1999	.610529	-.8662681	1.464436	-.3873031	-.3894081
9.	2000	1.573692	-1.478257	-1.301894	-.2857831	-.0266602

图 8.34　通过命令形式实现

图 8.35 展示的是系统提取的 5 个主因子的相关系数矩阵。



```
. correlate f1 f2 f3 f4 f5
(obs=9)
```

	f1	f2	f3	f4	f5
f1	1.0000				
f2	-0.0114	1.0000			
f3	0.0539	0.1023	1.0000		
f4	-0.0158	0.0218	0.1299	1.0000	
f5	0.0570	0.4102	-0.0757	-0.0517	1.0000

图 8.35　5 个主因子的相关系统矩阵

从图 8.35 中可以看出，我们提取的 5 个主因子之间几乎没有什么相关关系，这也说明了我们在前面对因子进行旋转的操作环节中采用最大方差正交旋转方式是明智的。

图 8.36 展示的是每个样本的因子得分示意图。

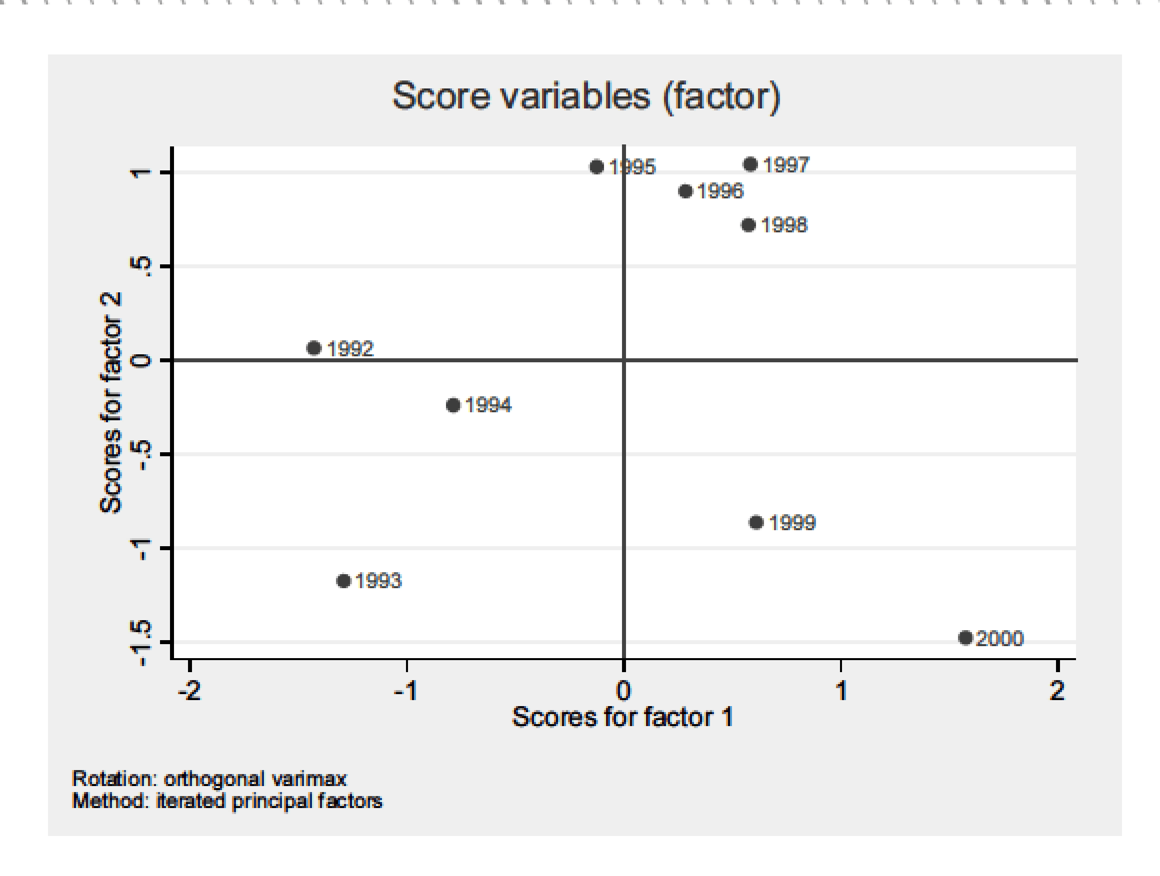


图 8.36　每个样本的因子得分示意图

从图 8.36 中可以看出，所有的样本被分到 4 个象限，其中第 1 象限包括 1996 年、1997 年、1998 年，这 3 年的两个因子得分都比较高；第 2 象限包括 1992 年、1995 年，这两年的因子 2 得分较高，而因子 1 得分较低；第 3 象限包括 1993 年、1994 年，这两年的两个因子得分都比较低；第 4 象限包括 1999 年、2000 年，这两年的因子 1 得分较高，而因子 2 得分较低。

图 8.37 展示的是本例因子分析的 KMO 检验结果。

```
. estat kmo

Kaiser-Meyer-Olkin measure of sampling adequacy
```

Variable	kmo
V2	0.6237
V3	0.6226
V4	0.7886
V5	0.1036
V6	0.6905
V7	0.7357
Overall	0.6566

图 8.37　KMO 检验结果

KMO 检验的结果与前面是一致的。

图 8.38 展示的是本例因子分析所提取的各个因子的特征值碎石图。

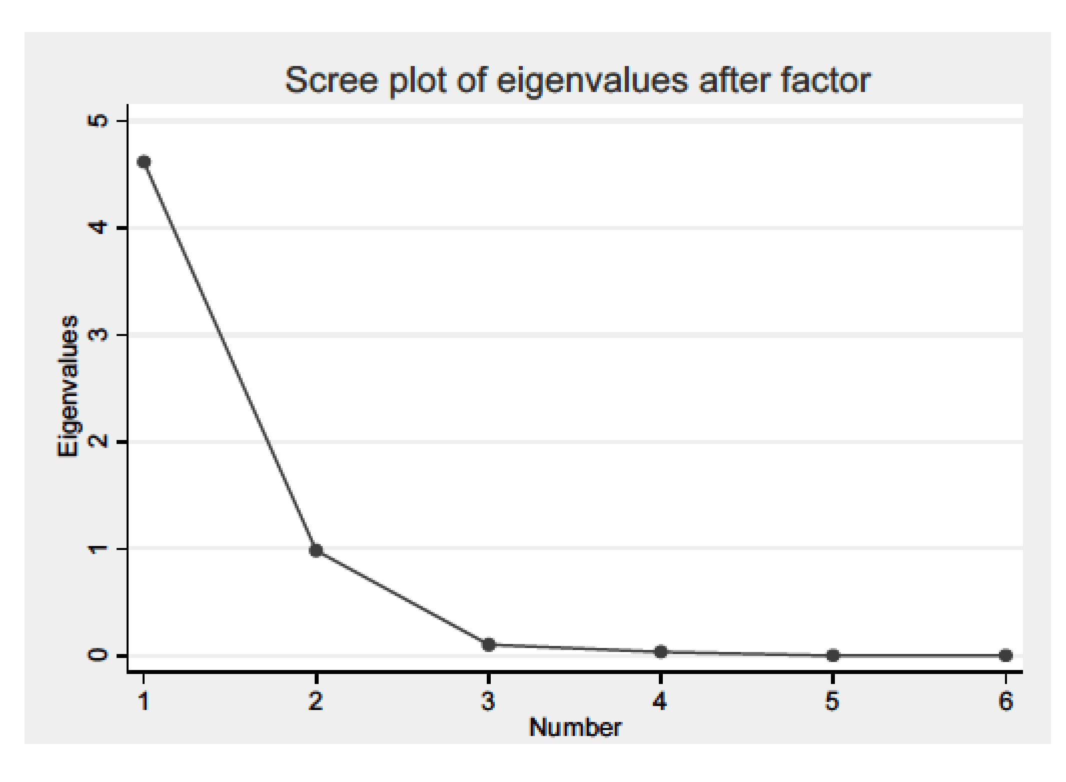
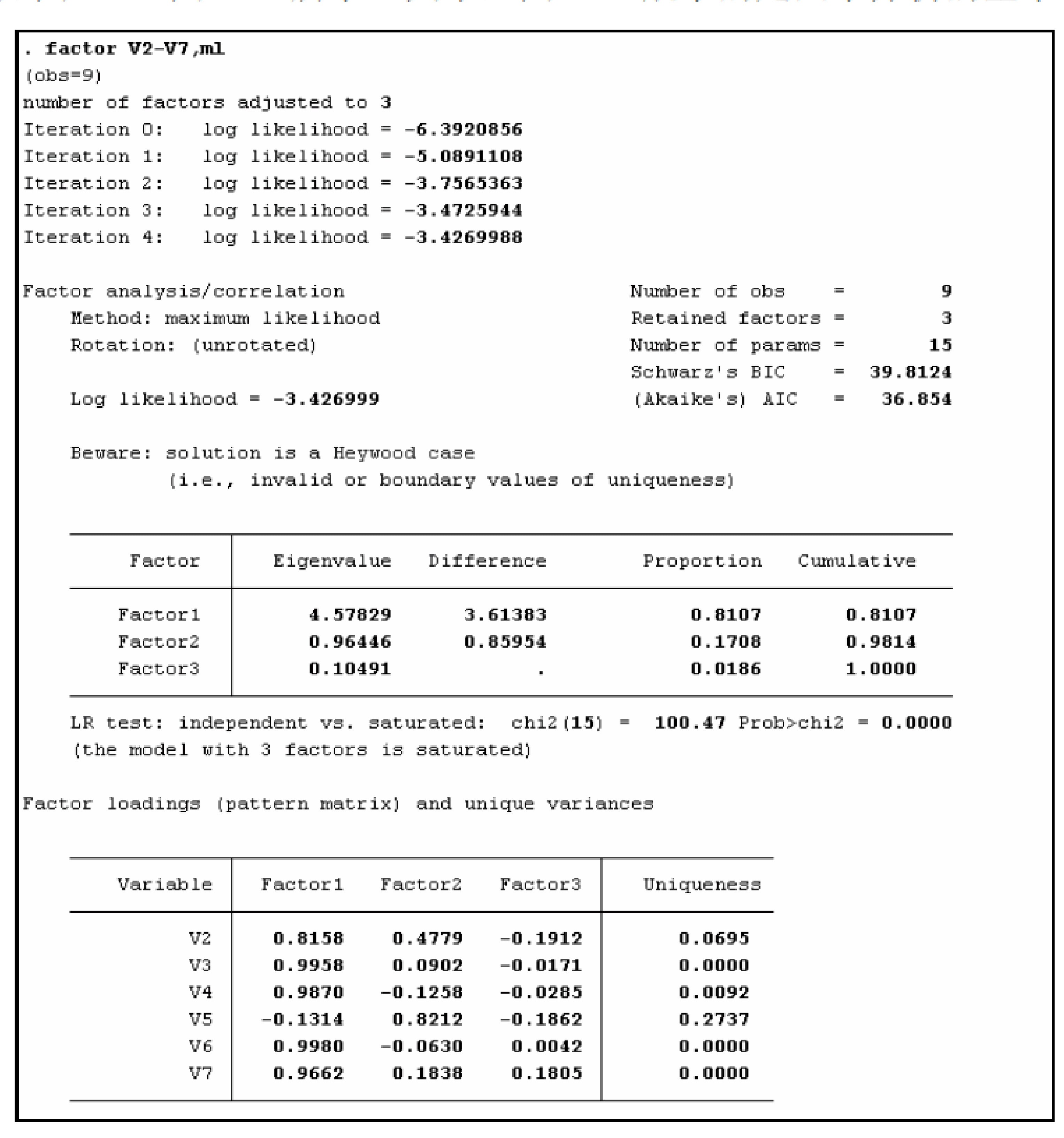


图 8.38　特征值碎石图

从图 8.38 中可以轻松地看出本例中只有第 1 个因子的特征值是明显大于 1 的，第 2 个因子的特征值是接近于 1 的。

4．最大似然因子法

分析结果如图 8.39~图 8.48 所示。其中，图 8.39 展示的是因子分析的基本情况。



```
. factor V2-V7,ml
(obs=9)
number of factors adjusted to 3
Iteration 0:   log likelihood = -6.3920856
Iteration 1:   log likelihood = -5.0891108
Iteration 2:   log likelihood = -3.7565363
Iteration 3:   log likelihood = -3.4725944
Iteration 4:   log likelihood = -3.4269988

Factor analysis/correlation                        Number of obs    =        9
    Method: maximum likelihood                     Retained factors =        3
    Rotation: (unrotated)                          Number of params =       15
                                                   Schwarz's BIC    =  39.8124
    Log likelihood = -3.426999                     (Akaike's) AIC   =   36.854

    Beware: solution is a Heywood case
            (i.e., invalid or boundary values of uniqueness)

    --------------------------------------------------------------------------
         Factor  |   Eigenvalue   Difference        Proportion   Cumulative
    -------------+------------------------------------------------------------
        Factor1  |      4.57829      3.61383            0.8107       0.8107
        Factor2  |      0.96446      0.85954            0.1708       0.9814
        Factor3  |      0.10491            .            0.0186       1.0000
    --------------------------------------------------------------------------
    LR test: independent vs. saturated:  chi2(15) =  100.47 Prob>chi2 = 0.0000
    (the model with 3 factors is saturated)

Factor loadings (pattern matrix) and unique variances

    -------------------------------------------------------------
        Variable |  Factor1   Factor2   Factor3 |   Uniqueness
    -------------+------------------------------+----------------
              V2 |   0.8158    0.4779   -0.1912 |      0.0695
              V3 |   0.9958    0.0902   -0.0171 |      0.0000
              V4 |   0.9870   -0.1258   -0.0285 |      0.0092
              V5 |  -0.1314    0.8212   -0.1862 |      0.2737
              V6 |   0.9980   -0.0630    0.0042 |      0.0000
              V7 |   0.9662    0.1838    0.1805 |      0.0000
    -------------------------------------------------------------
```

图 8.39　因子分析的基本情况

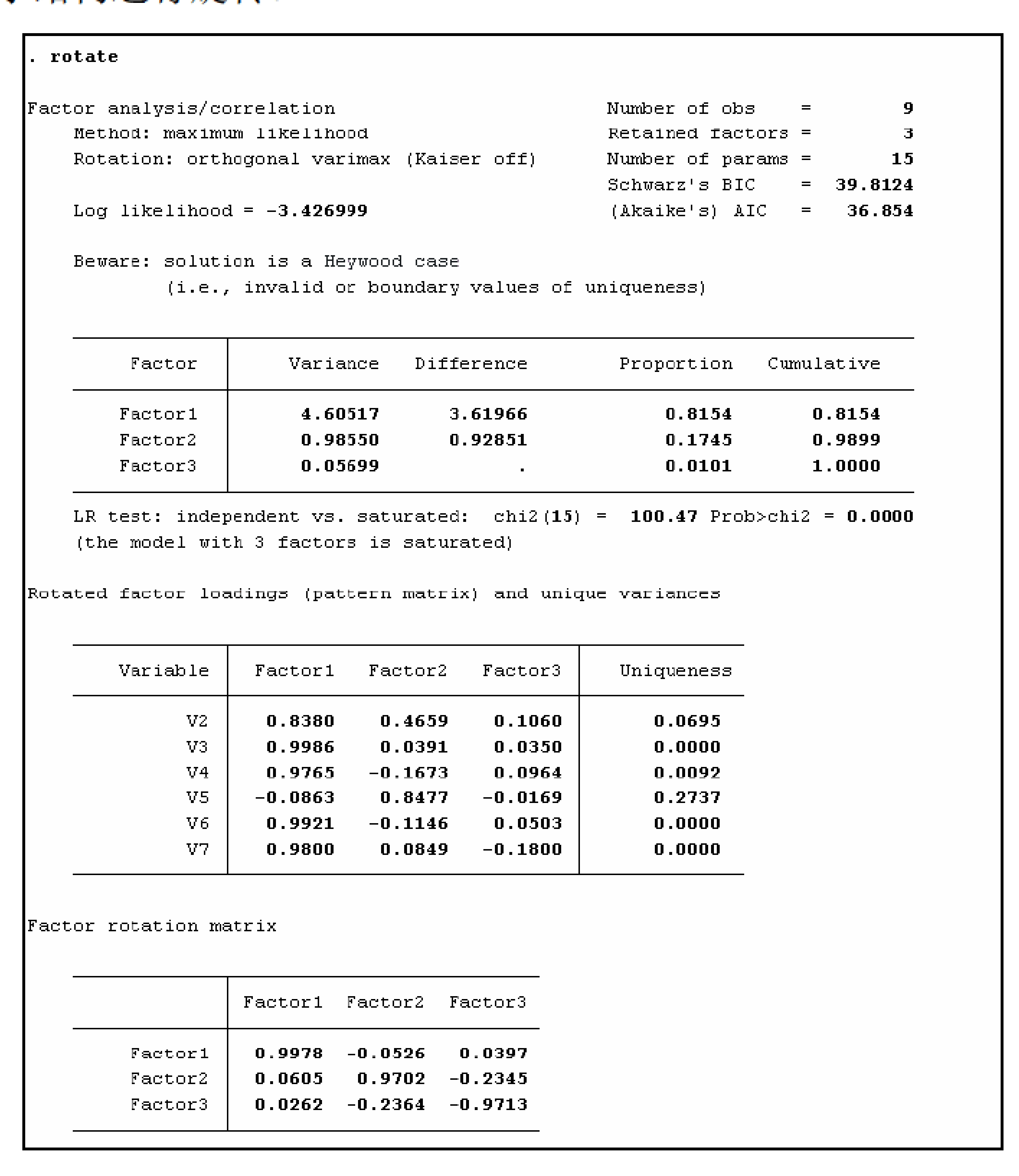
该检验有助于确定合适的因子数目。图 8.39 的第 1 部分说明的是因子分析经过迭代计算后在第 4 次（Iteration 4: log likelihood = -3.4269988）达到饱和，此时系统提取的主因子个数是 3 个。

从图 8.39 的第 2 部分我们可以看出共有 9 个样本（Number of obs= 9）参与了分析，BIC 信息准则值为 39.8124，AIC 信息准则值为 36.854，模型 LR 检验的卡方值（LR test: independent vs. saturated: chi2(15)）为 100.47，P 值（Prob>chi2）为 0.0000，模型非常显著。图 8.39 的第 2 部分最左列（Factor）说明的是因子名称。Eigenvalue 列表示的是提取因子的特征值情况，只有第 1 个因子的特征值是大于 1 的，其中第 1 个因子的特征值是 4.57829，第 2 个因子的特征值是 0.96446。Proportion 列表示的是提取因子的方差贡献率，其中第 1 个因子的方差贡献率为 81.07%，第 2 个因子的方差贡献率为 17.08%。Cumulative 列表示的是提取因子的累计方差贡献率，其中前两个因子的累计方差贡献率为 98.14%。

图 8.39 的下半部分说明的是模型的因子载荷矩阵以及变量的未被解释部分。其中，Variable 列表示的是变量名称，Factor1、Factor2、Factor3 这 3 列分别说明的是提取的 3 个主因子对各

个变量的解释程度，本例中，Factor1 主要解释的是 V2、V3、V4、V6、V7 这 5 个变量的信息，Factor2 主要解释的是 V5 变量的信息。Uniqueness 列表示变量未被提取的前两个主因子解释的部分，可以发现在舍弃其他主因子的情况下，信息的损失量是很小的。

图 8.40 展示的是对因子结构进行旋转的结果。此处我们依然采用系统默认的最大方差正交旋转方式对因子结构进行旋转。



```
. rotate

Factor analysis/correlation                        Number of obs    =          9
    Method: maximum likelihood                     Retained factors =          3
    Rotation: orthogonal varimax (Kaiser off)      Number of params =         15
                                                   Schwarz's BIC    =    39.8124
    Log likelihood = -3.426999                     (Akaike's) AIC   =     36.854

    Beware: solution is a Heywood case
            (i.e., invalid or boundary values of uniqueness)

    --------------------------------------------------------------------------
         Factor  |     Variance   Difference        Proportion   Cumulative
    -------------+------------------------------------------------------------
        Factor1  |      4.60517      3.61966            0.8154       0.8154
        Factor2  |      0.98550      0.92851            0.1745       0.9899
        Factor3  |      0.05699            .            0.0101       1.0000
    --------------------------------------------------------------------------
    LR test: independent vs. saturated:  chi2(15) =  100.47 Prob>chi2 = 0.0000
    (the model with 3 factors is saturated)

Rotated factor loadings (pattern matrix) and unique variances

    ---------------------------------------------------------
        Variable |  Factor1   Factor2   Factor3 |   Uniqueness
    -------------+------------------------------+--------------
              V2 |   0.8380    0.4659    0.1060 |       0.0695
              V3 |   0.9986    0.0391    0.0350 |       0.0000
              V4 |   0.9765   -0.1673    0.0964 |       0.0092
              V5 |  -0.0863    0.8477   -0.0169 |       0.2737
              V6 |   0.9921   -0.1146    0.0503 |       0.0000
              V7 |   0.9800    0.0849   -0.1800 |       0.0000
    ---------------------------------------------------------

Factor rotation matrix

    --------------------------------------------
                 |  Factor1  Factor2  Factor3
    -------------+------------------------------
         Factor1 |   0.9978  -0.0526   0.0397
         Factor2 |   0.0605   0.9702  -0.2345
         Factor3 |   0.0262  -0.2364  -0.9713
    --------------------------------------------
```

图 8.40　对因子结构进行旋转

图 8.40 包括 3 部分内容，第 1 部分说明的是因子旋转模型的一般情况，从图中我们可以看出共有 9 个样本（Number of obs = 9）参与了分析，提取保留的因子共有 3 个（Retained factors = 3），模型 LR 检验的卡方值（LR test: independent vs. saturated: chi2(15)）为 100.47，P 值（Prob>chi2）为 0.0000，模型非常显著。最左列（Factor）说明的是因子名称，可以看出模型旋转后共提取了 3 个因子。Proportion 列表示的是提取因子的方差贡献率，其中第 1 个因子的方差贡献率为 81.54%，第 2 个因子的方差贡献率为 17.45%。Cumulative 列表示的是提取因子的累计方差贡献率，其中前两个因子的累计方差贡献率为 98.99%。

图 8.40 的第 2 部分说明的是模型的因子载荷矩阵以及变量的未被解释部分。其中，Variable 列表示的是变量名称，Factor1、Factor2、Factor3 这 3 列分别说明的是旋转提取的 3 个主因子对各个变量的解释程度，本例中，Factor1 主要解释的是 V2、V3、V4、V6、V7 这 5 个变量的信息，Factor2 主要解释的是 V5 变量的信息。Uniqueness 列表示变量未被提取的前两个主因子解释的部分，可以发现在舍弃其他主因子的情况下，信息的损失量是很小的。

图 8.40 的第 3 部分展示的是因子旋转矩阵的一般情况，提取的 3 个因子相关关系很弱。

图 8.41 展示的因子旋转后的因子载荷图。此处我们通过 Factor 选项控制了因子的数目，本因子载荷图可以使用户更加直观地看出各个变量被前两个因子的解释情况。

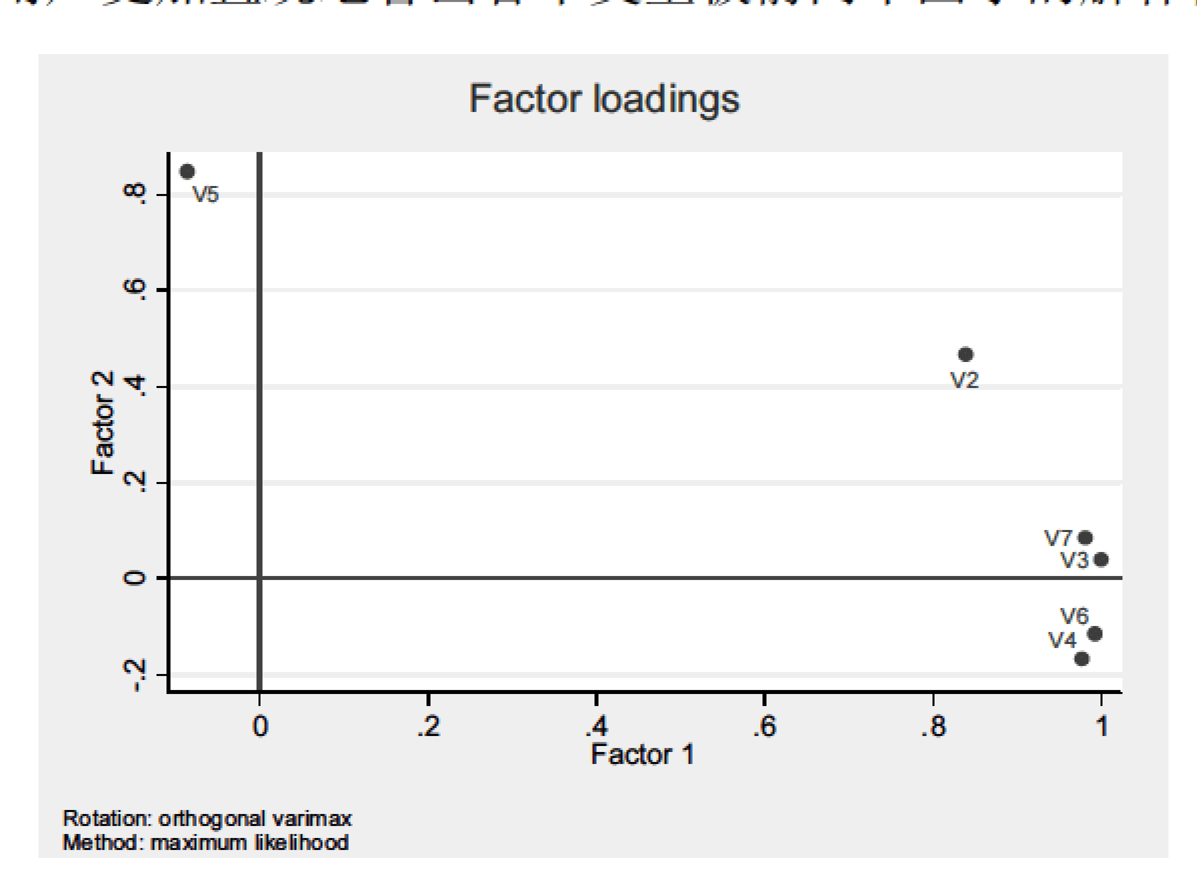


图 8.41　旋转后的因子载荷图

与前面的分析相同，V2、V3、V4、V6、V7 这 5 个变量的信息主要被 Factor1 这一因子所解释，V5 变量主要被 Factor2 这一因子所解释。

图 8.42 展示的是因子分析后各个样本的因子得分情况。

```
. predict f1 f2 f3
(regression scoring assumed)

Scoring coefficients (method = regression; based on varimax rotated factors)
```

Variable	Factor1	Factor2	Factor3
V2	0.00001	0.00001	0.00001
V3	0.50900	7.09107	6.11430
V4	0.00002	-0.00000	0.00002
V5	0.00000	0.00000	0.00000
V6	0.31163	-6.66082	-1.44353
V7	0.18623	-0.48252	-4.76910

图 8.42　各个样本的因子得分情况

根据图 8.42 展示的因子得分系数矩阵，可以写出各公因子的表达式。值得一提的是，在表达式中各个变量已经不是原始变量，而是标准化变量。

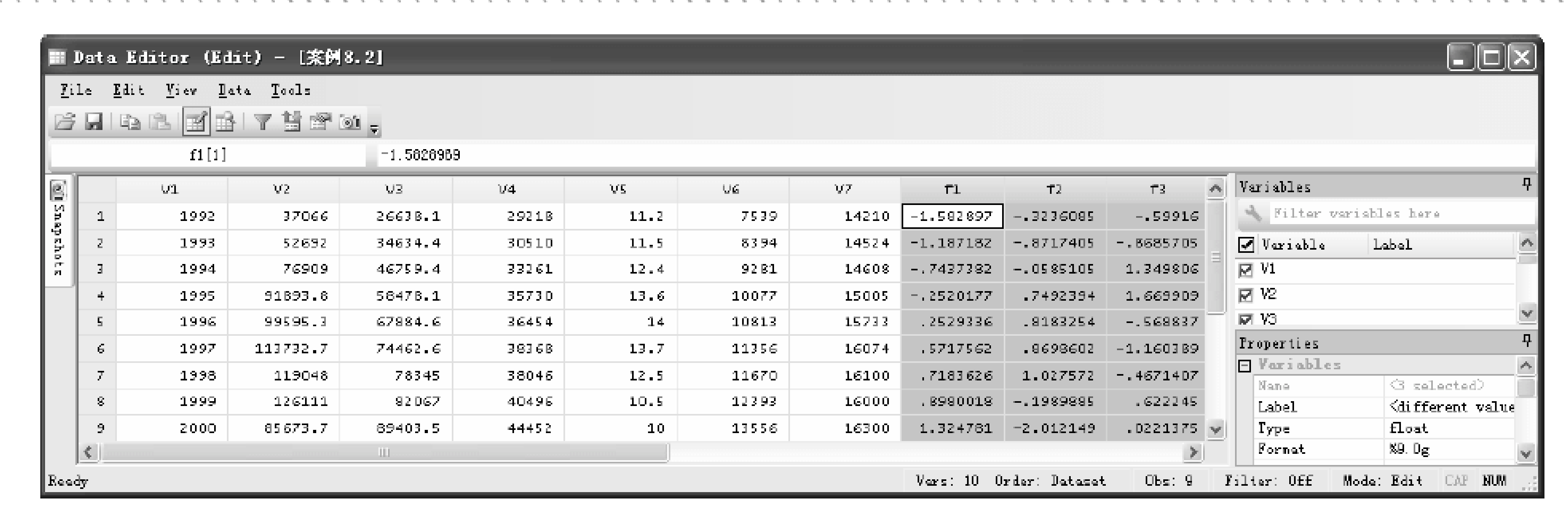
表达式如下：

F1=0.00001*工业总产值+0.50900*国内生产总值+ 0.00002*货物周转量+0.00000*原煤+0.31163*发电量+0.18623*原油

F2=0.00001*工业总产值+7.09107*国内生产总值-0.00000*货物周转量+0.00000*原煤-6.66082*发电量-0.48252*原油

F3=0.00001*工业总产值+6.11430*国内生产总值+0.00002*货物周转量+0.00000*原煤-1.44353*发电量-4.76910*原油

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 8.43 所示的因子得分数据。



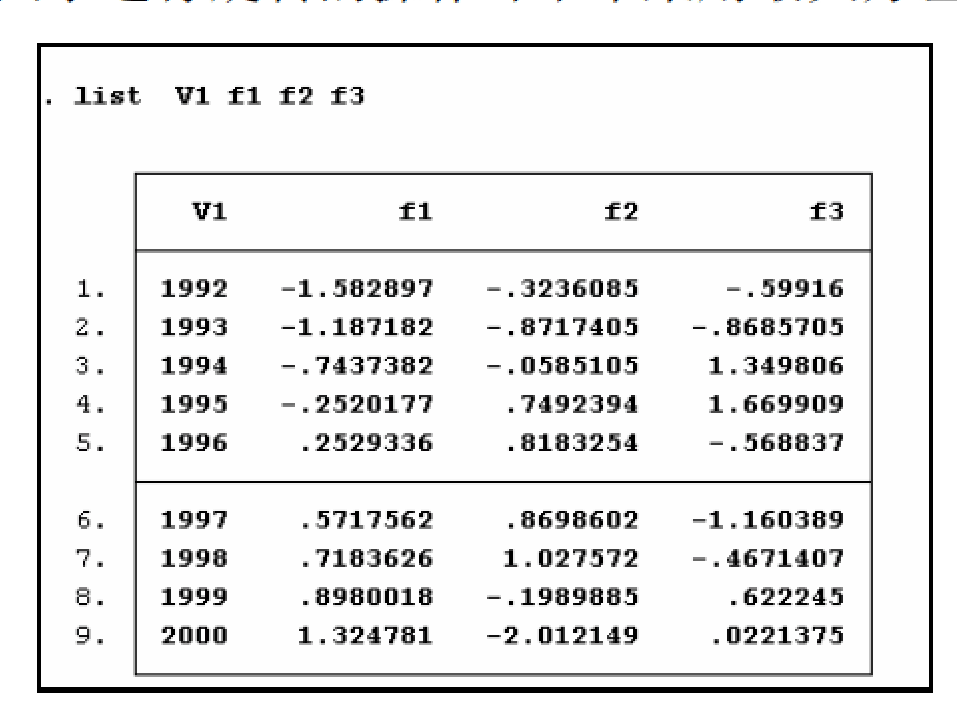
	V1	V2	V3	V4	V5	V6	V7	f1	f2	f3
1	1992	37066	26638.1	29218	11.2	7539	14210	-1.582897	-.3236085	-.59916
2	1993	52692	34634.4	30510	11.5	8394	14524	-1.187182	-.8717405	-.8685705
3	1994	76909	46759.4	33261	12.4	9281	14608	-.7437382	-.0585105	1.349806
4	1995	91893.8	58478.1	35730	13.6	10077	15005	-.2520177	.7492394	1.669909
5	1996	99595.3	67884.6	36454	14	10813	15733	.2529336	.8183254	-.568837
6	1997	113732.7	74462.6	38368	13.7	11356	16074	.5717562	.8698602	-1.160389
7	1998	119048	78345	38046	12.5	11670	16100	.7183626	1.027572	-.4671407
8	1999	126111	82067	40496	10.5	12393	16000	.8980018	-.1989885	.622245
9	2000	85673.7	89403.5	44452	10	13556	16300	1.324781	-2.012149	.0221375

图 8.43　数据查看界面

这一点也可以通过命令形式实现，如图 8.44 所示。

图 8.45 展示的是系统提取的 3 个主因子的相关系数矩阵。

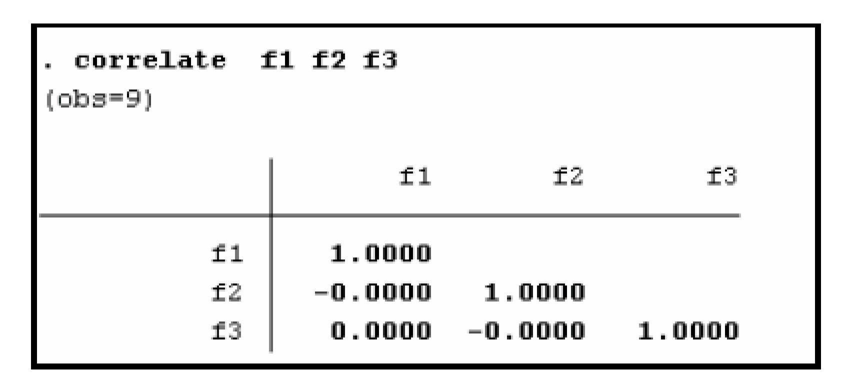
从图 8.45 中可以看出，提取的 3 个主因子之间几乎没有什么相关关系，这也说明了在前面对因子进行旋转的操作环节中采用最大方差正交旋转方式是明智的。



```
. list  V1 f1 f2 f3
```

	V1	f1	f2	f3
1.	1992	-1.582897	-.3236085	-.59916
2.	1993	-1.187182	-.8717405	-.8685705
3.	1994	-.7437382	-.0585105	1.349806
4.	1995	-.2520177	.7492394	1.669909
5.	1996	.2529336	.8183254	-.568837
6.	1997	.5717562	.8698602	-1.160389
7.	1998	.7183626	1.027572	-.4671407
8.	1999	.8980018	-.1989885	.622245
9.	2000	1.324781	-2.012149	.0221375

图 8.44　通过命令形式实现



```
. correlate  f1 f2 f3
(obs=9)
```

	f1	f2	f3
f1	1.0000		
f2	-0.0000	1.0000	
f3	0.0000	-0.0000	1.0000

图 8.45　3 个主因子的相关系统矩阵

图 8.46 展示的是每个样本的因子得分示意图。

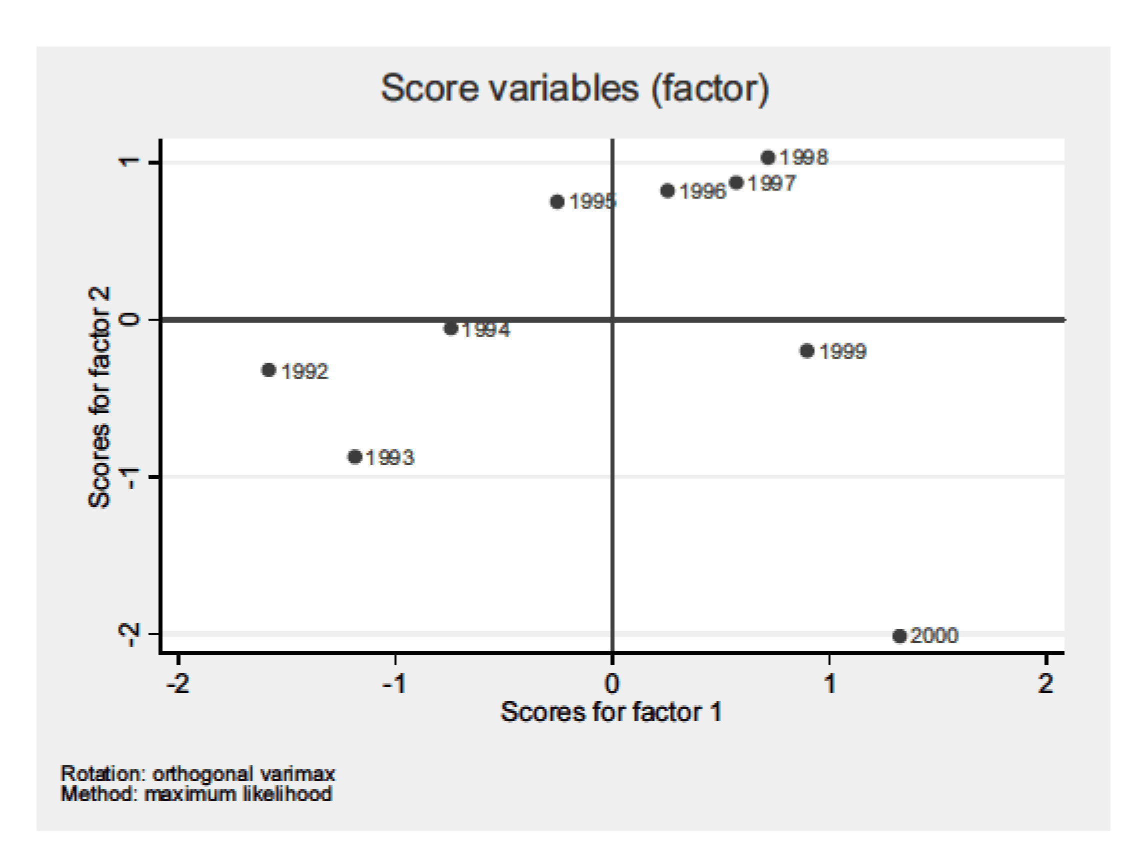


图 8.46　每个样本的因子得分示意图

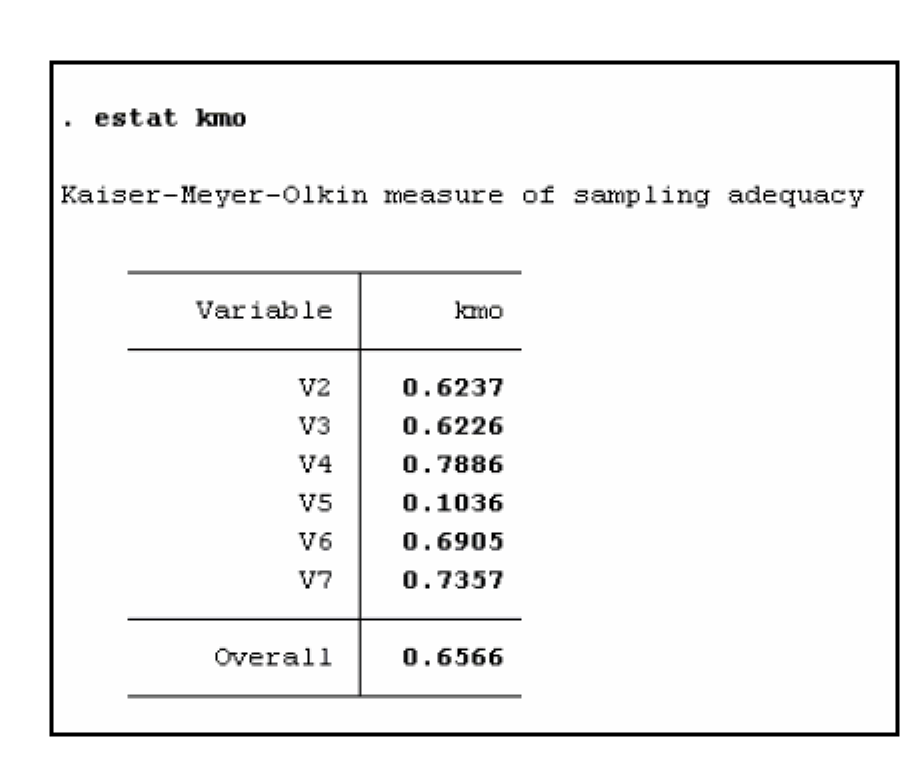
从图 8.46 中可以看出，所有的样本被分到 4 个象限，其中第 1 象限包括 1996 年、1997 年、1998 年，这 3 年的两个因子得分都比较高；第 2 象限包括 1995 年，这一年的因子 2 得分较高，而因子 1 得分较低；第 3 象限包括 1992 年、1993 年、1994 年，这 3 年的两个因子得分

都比较低；第 4 象限包括 1999 年、2000 年，这两年的因子 1 得分较高，而因子 2 得分较低。

图 8.47 展示的是本例因子分析的 KMO 检验结果。

KMO 检验的结果与前面是一致的。

图 8.48 展示的是本例因子分析所提取的各个因子的特征值碎石图。



```
. estat kmo

Kaiser-Meyer-Olkin measure of sampling adequacy
```

Variable	kmo
V2	0.6237
V3	0.6226
V4	0.7886
V5	0.1036
V6	0.6905
V7	0.7357
Overall	0.6566



图 8.47　KMO 检验结果

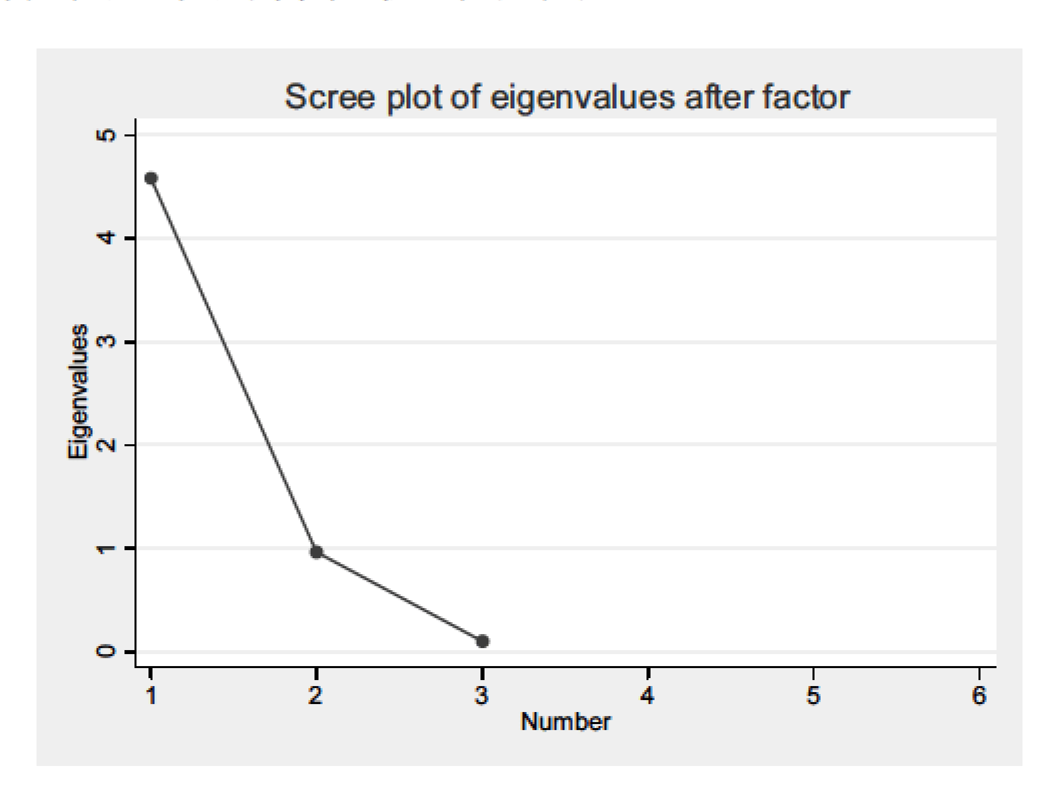


图 8.48　特征值碎石图

从图 8.48 中可以轻松地看出本例中只有第 1 个因子的特征值是明显大于 1 的，第 2 个因子的特征值是接近于 1 的。

8.2.5　案例延伸

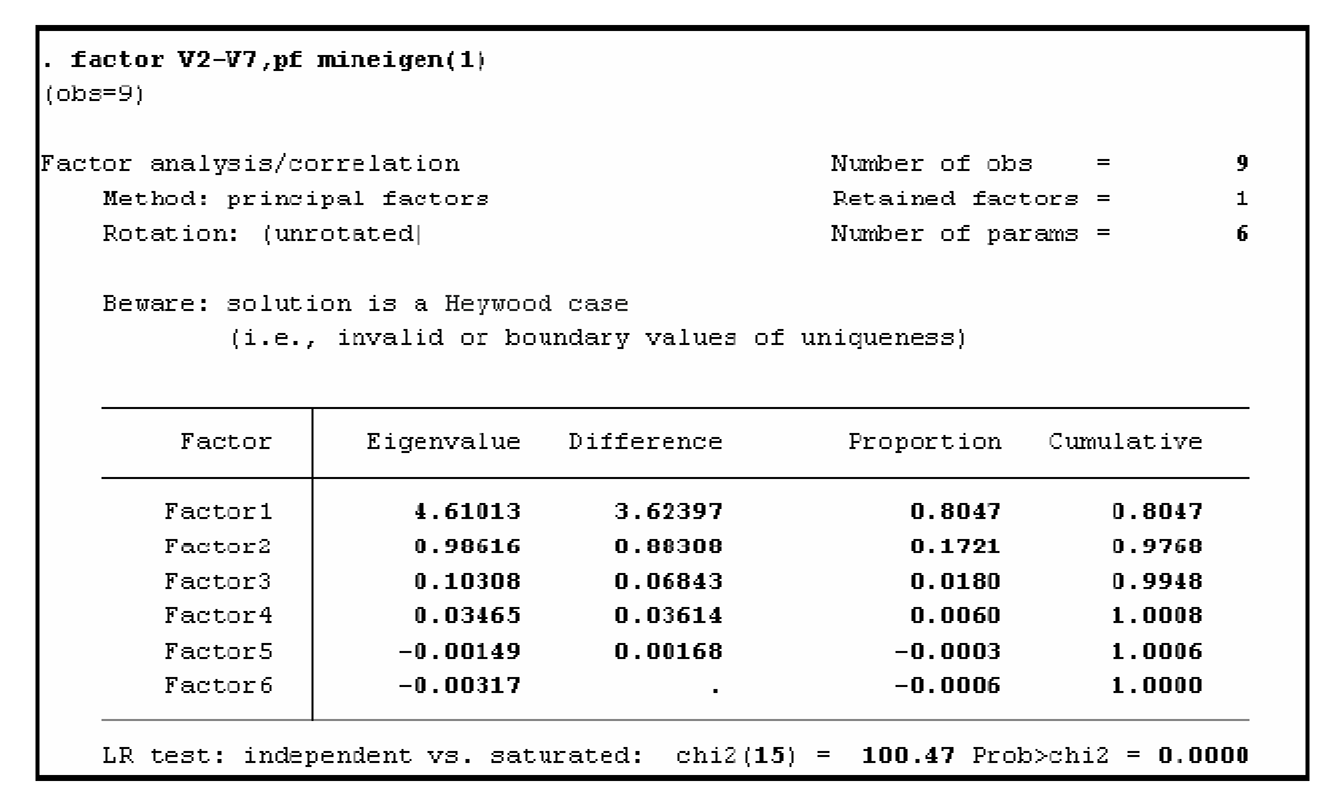
上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

1．延伸 1：只保留特征值大于一定值的操作选项

例如，在本节例子的主成分因子法操作中，我们只保留特征值大于 1 的因子，操作命令应该相应地修改为：

```
factor V2-V7,pf mineigen(1)
```

在命令窗口输入命令并按回车键进行确认，结果如图 8.49~图 8.50 所示。



```
. factor V2-V7,pf mineigen(1)
(obs=9)

Factor analysis/correlation                   Number of obs    =        9
    Method: principal factors                 Retained factors =        1
    Rotation: (unrotated)                     Number of params =        6

    Beware: solution is a Heywood case
            (i.e., invalid or boundary values of uniqueness)
```

Factor	Eigenvalue	Difference	Proportion	Cumulative
Factor1	4.61013	3.62397	0.8047	0.8047
Factor2	0.98616	0.88308	0.1721	0.9768
Factor3	0.10308	0.06843	0.0180	0.9948
Factor4	0.03465	0.03614	0.0060	1.0008
Factor5	-0.00149	0.00168	-0.0003	1.0006
Factor6	-0.00317	.	-0.0006	1.0000

```
    LR test: independent vs. saturated:  chi2(15) =  100.47 Prob>chi2 = 0.0000
```

图 8.49　分析结果图 1

图 8.50 展示的内容与结果分析部分所展示的是一致的。

Factor loadings (pattern matrix) and unique variances

Variable	Factor1	Uniqueness
V2	0.8628	0.2556
V3	1.0001	-0.0002
V4	0.9682	0.0626
V5	-0.0587	0.9966
V6	0.9872	0.0254
V7	0.9747	0.0499

图 8.50　分析结果图 2

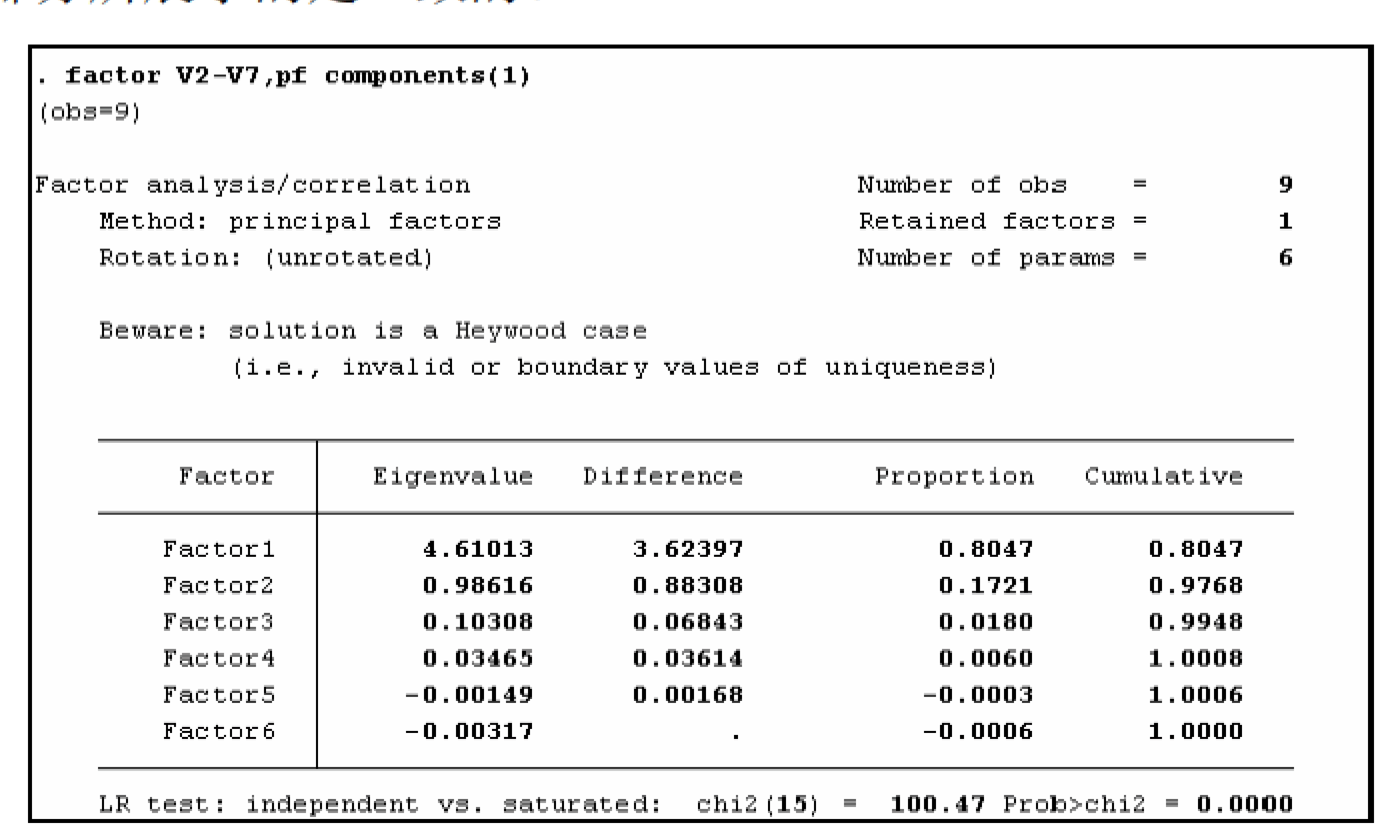
图 8.50 展示的是仅保留特征值大于 1 的主成分的结果，本例中只有 1 个主成分的特征值是大于 1 的，所以只保留了 1 个主成分进行分析。Uniqueness 列表示变量未被提取的主成分解释的部分，例如变量 V2 未被解释的信息比例就是 25.56%。这种信息丢失情况是我们舍弃其他主成分必然付出的代价。

2．延伸 2：限定提取的主成分个数的操作选项

例如，在本节例子的主成分因子法操作中，我们只想提取一个主成分进行分析，那么操作命令应该相应地修改为：

```
factor V2-V7,pf components(1)
```

在命令窗口输入命令并按回车键进行确认，结果如图 8.51~图 8.52 所示。图 8.51 展示的内容与结果分析部分所展示的是一致的。



```
. factor V2-V7,pf components(1)
(obs=9)

Factor analysis/correlation                  Number of obs    =      9
    Method: principal factors                Retained factors =      1
    Rotation: (unrotated)                    Number of params =      6

    Beware: solution is a Heywood case
            (i.e., invalid or boundary values of uniqueness)
```

Factor	Eigenvalue	Difference	Proportion	Cumulative
Factor1	4.61013	3.62397	0.8047	0.8047
Factor2	0.98616	0.88308	0.1721	0.9768
Factor3	0.10308	0.06843	0.0180	0.9948
Factor4	0.03465	0.03614	0.0060	1.0008
Factor5	-0.00149	0.00168	-0.0003	1.0006
Factor6	-0.00317	.	-0.0006	1.0000

LR test: independent vs. saturated: chi2(15) = 100.47 Prob>chi2 = 0.0000

图 8.51　分析结果图 1

Factor loadings (pattern matrix) and unique variances

Variable	Factor1	Uniqueness
V2	0.8628	0.2556
V3	1.0001	-0.0002
V4	0.9682	0.0626
V5	-0.0587	0.9966
V6	0.9872	0.0254
V7	0.9747	0.0499

图 8.52　分析结果图 2

图 8.52 展示的是我们只提取一个主成分进行分析的结果，该图最后一列（Uniqueness）同样说明的是该变量未被系统提取的一个主成分解释的信息比例，例如变量 V2 未被解释的信息比例就是 25.56%。这种信息丢失情况同样也是我们舍弃其他主成分必然付出的代价。

8.3 本章习题

（1）表 8.3 给出了我国历年国民经济主要指标统计数据（1996－2003 年）。试对这些指标进行主成分分析。

表 8.3 我国历年国民经济主要指标统计数据（1996－2003 年）

年份	工业总产值/亿元	国内生产总值/亿元	货物周转量/亿吨千米	原煤/亿吨	发电量/亿千瓦时	原油/万吨
1996	99595.3	67884.6	36590.0	14.0	10813.0	15733.0
1997	113732.7	74462.6	38385.0	13.7	11356.0	16074.0
1998	119048.0	78345.0	38089.0	12.5	11670.0	16100.0
1999	126111.0	82067.0	40568.0	10.5	12393.0	16000.0
2000	85673.7	89442.0	44321.0	10.0	13556.0	16300.0
2001	95449.0	97315.0	47710.0	11.6	14808.0	16396.0
2002	110776.0	105172.0	50686.0	13.8	16540.0	16700.0
2003	142271.0	117251.9	53859.0	16.7	19106.0	16960.0

（2）对表 8.3 所给出的资料进行因子分析。

第 9 章　Stata 聚类分析

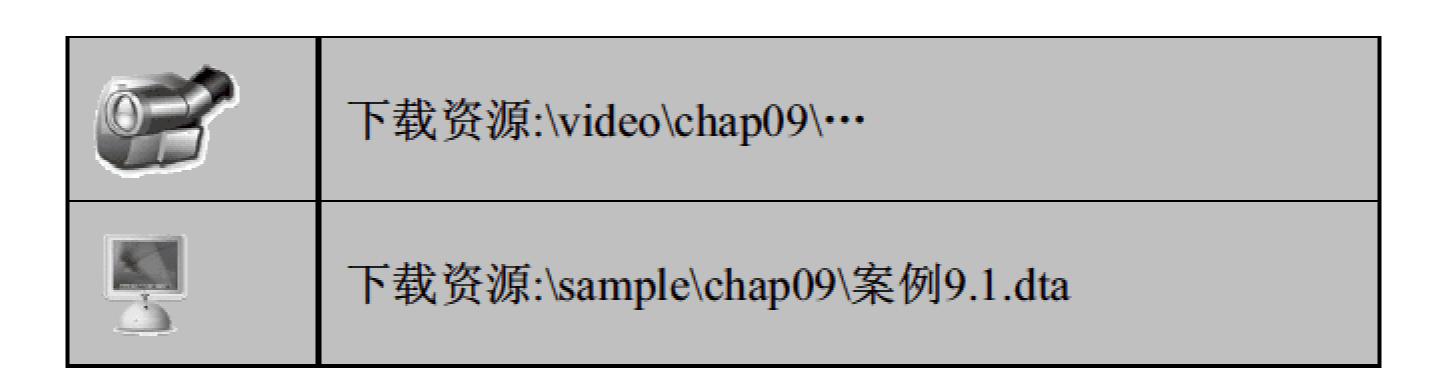
聚类分析（Cluster Analysis）是研究事物分类的基本方法，基于我们所研究的指标或数据之间存在着的不同程度的相似性或者相异性。聚类分析采用定量数学方法，根据样品或指标的数值特征对样品进行分类，从而辨别出各样品之间的亲疏关系。聚类分析是一种使用简单但却很常用的分析方法，往往被用来进行经验性类型的探索，而不是用来检验事先所定的假设。聚类分析分成两个宽泛的类别，包括划分聚类分析和层次聚类分析。本章将逐一介绍这两种聚类分析方法在实例中的应用。

9.1　实例一——划分聚类分析

9.1.1　划分聚类分析的功能与意义

划分聚类分析方法（Partition）的基本思想是将观测到的样本划分到一系列事先设定好的不重合的分组中去。划分聚类分析方法在计算上相比层次聚类分析方法要相对简单而且计算速度要更快一些，但是它也有自己的缺点，它要求事先指定样本聚类的精确数目，这与聚类分析探索性的本质是不相适应的。划分聚类分析包括两种：一种是 K 个平均数的聚类分析方法（Cluster Kmeans），此方法的操作流程是通过迭代过程将观测案例分配到具有最接近的平均数的组，然后找出这些聚类；另一种是 K 个中位数的聚类分析方法（Cluster Kmedians），此方法的操作流程是通过迭代过程将观测案例分配到具有最接近的中位数的组，然后找出这些聚类。下面我们就以实例的方式介绍一下这两种划分聚类分析方法。

9.1.2　相关数据来源



	下载资源:\video\chap09\…
	下载资源:\sample\chap09\案例9.1.dta

【例 9.1】表 9.1 是我国 2006 年各地区能源消耗的情况。根据不同省市的能源消耗情况，对其进行划分聚类分析，以便了解我国不同地区的能源消耗情况。

表 9.1　2006 年各地区能源消耗统计表

地区	单位地区生产总值煤消耗量/吨	单位地区生产总值电消耗量/千瓦/时	单位工业增加值煤消耗量/吨
北京	0.8	828.5	1.5
天津	1.11	1040.8	1.45
河北	1.96	1487.6	4.41
山西	2.95	2264.2	6.57
内蒙古	2.48	1714.1	5.67
…	…	…	…
青海	3.07	3801.8	3.44
宁夏	4.14	4997.7	9.03
新疆	2.11	1190.9	3.00

9.1.3　Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有 4 个变量，分别是地区、单位地区生产总值煤消耗量（吨）、单位地区生产总值电消耗量（千瓦/时）、单位工业增加值煤消耗量（吨）。我们把这些变量分别定义为 V1、V2、V3、V4，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作我们在第 1 章中已有详细讲述。录入完成后数据如图 9.1 所示。

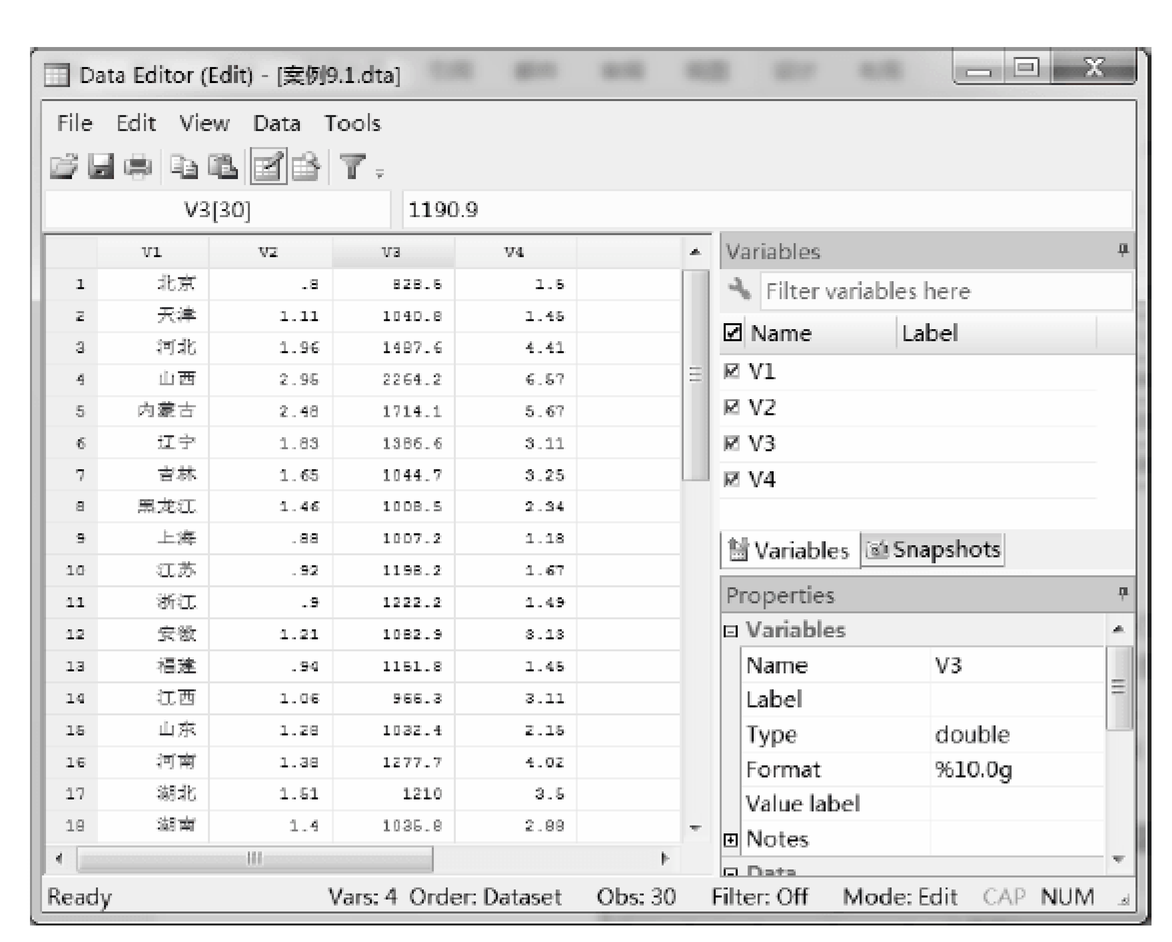


图 9.1　案例 9.1 数据

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中分别输入如下命令并按键盘上的回车键进行确认。

- egen zv2=std(V2)：本命令旨在对 V2 变量进行标准化处理。
- egen zv3=std(V3)：本命令旨在对 V3 变量进行标准化处理。
- egen zv4=std(V4)：本命令旨在对 V4 变量进行标准化处理。
- sum zv2 zv3 zv4：本命令旨在对 zv2、zv3、zv4 变量进行描述性统计分析。

- cluster kmeans zv2 zv3 zv4,k(2)：本命令的含义是对 zv2、zv3、zv4 变量进行 K 个平均数的聚类分析，并把样本分为 2 类。
- cluster kmeans zv2 zv3 zv4,k(3)：本命令的含义是对 zv2、zv3、zv4 变量进行 K 个平均数的聚类分析，并把样本分为 3 类。
- cluster kmeans zv2 zv3 zv4,k(4)：本命令的含义是对 zv2、zv3、zv4 变量进行 K 个平均数的聚类分析，并把样本分为 4 类。
- cluster kmedians zv2 zv3 zv4,k(2)：本命令的含义是对 zv2、zv3、zv4 变量进行 K 个中位数的聚类分析，并把样本分为 2 类。
- cluster kmedians zv2 zv3 zv4,k(3)：本命令的含义是对 zv2、zv3、zv4 变量进行 K 个中位数的聚类分析，并把样本分为 3 类。
- cluster kmedians zv2 zv3 zv4,k(4)：本命令的含义是对 zv2、zv3、zv4 变量进行 K 个中位数的聚类分析，并把样本分为 4 类。

03 设置完毕后，按键盘上的回车键，等待输出结果。

9.1.4 结果分析

在 Stata 14.0 主界面的结果窗口我们可以看到如图 9.2~图 9.17 所示的分析结果。

1．数据标准化处理

在分析过程中前 3 条 Stata 命令旨在对数据进行标准化处理，选择的标准化处理方式是使变量的平均数为 0 而且标准差为 1。之所以这样做是因为我们进行聚类分析的变量都是以不可比的单位进行的测度，它们具有极为不同的方差，我们对数据进行标准化处理可以避免使结果受到具有最大方差变量的影响。在输入前 3 条 Stata 命令并且分别按键盘上的回车键进行确认后，选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 9.2 所示的变换后的数据。

根据我们在前面章节中讲述的描述性统计分析方法，我们可以看到如图 9.3 所示的标准化变量的相应统计量。

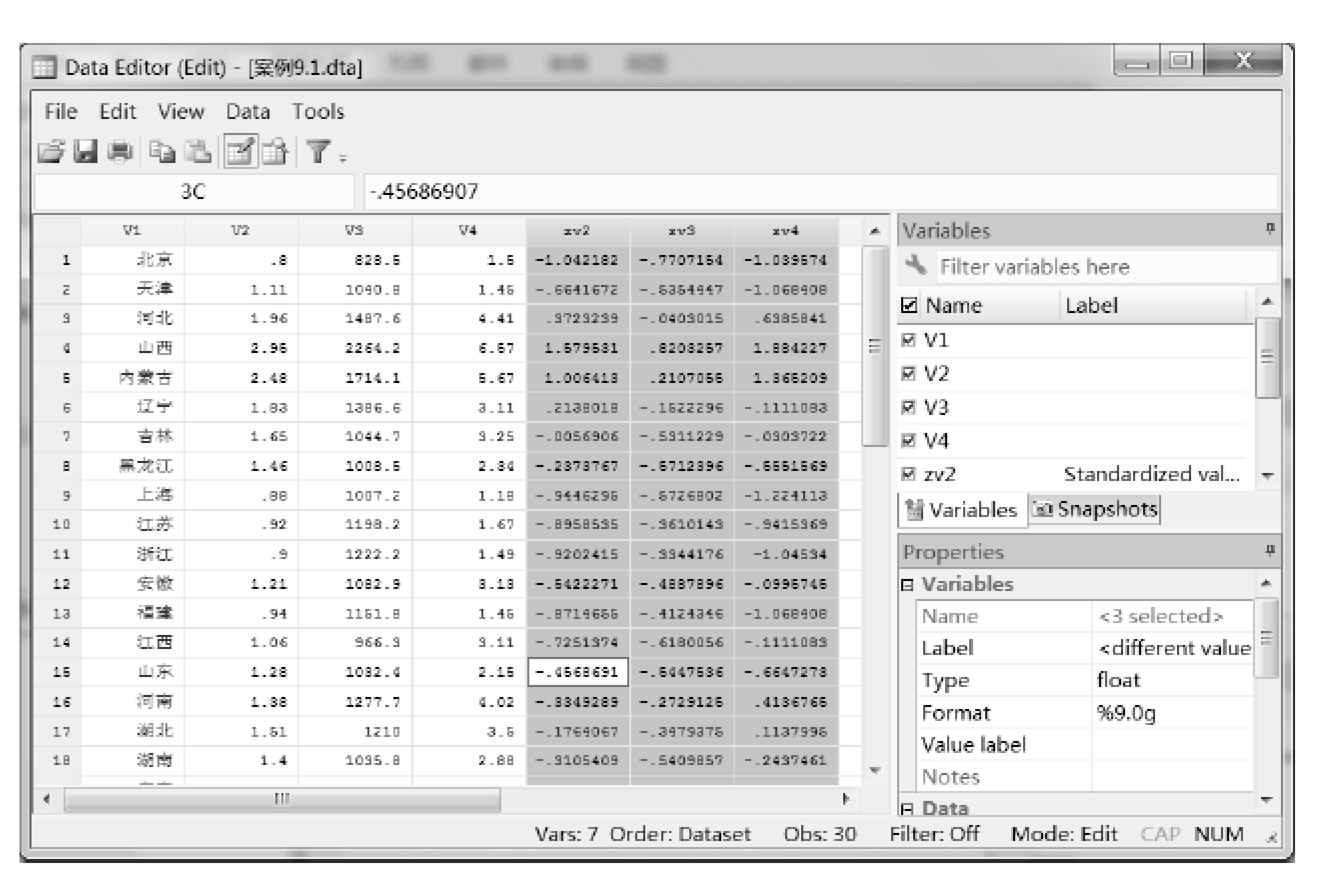


图 9.2　标准化变换后的数据

```
. summ  zv2 zv3 zv4

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         zv2 |        30    7.67e-09           1  -1.054376   3.030619
         zv3 |        30    7.70e-09           1  -.7707154   3.849588
         zv4 |        30   -5.77e-09           1  -1.281782   3.302876
```

图 9.3　标准化变量的相应统计量

通过观察分析结果，我们可以看出，有效观测样本共有 30 个。zv2 的平均值为 7.67e-09，标准差是 1，最小值是-1.054376，最大值是 3.030619；zv3 的平均值为 7.70e-09，标准差是 1，最小值是-0.7707154，最大值是 3.849588；zv4 的平均值为-5.77e-09，标准差是 1，最小值是-1.281782，最大值是 3.302876。

2．K 个平均数的聚类分析

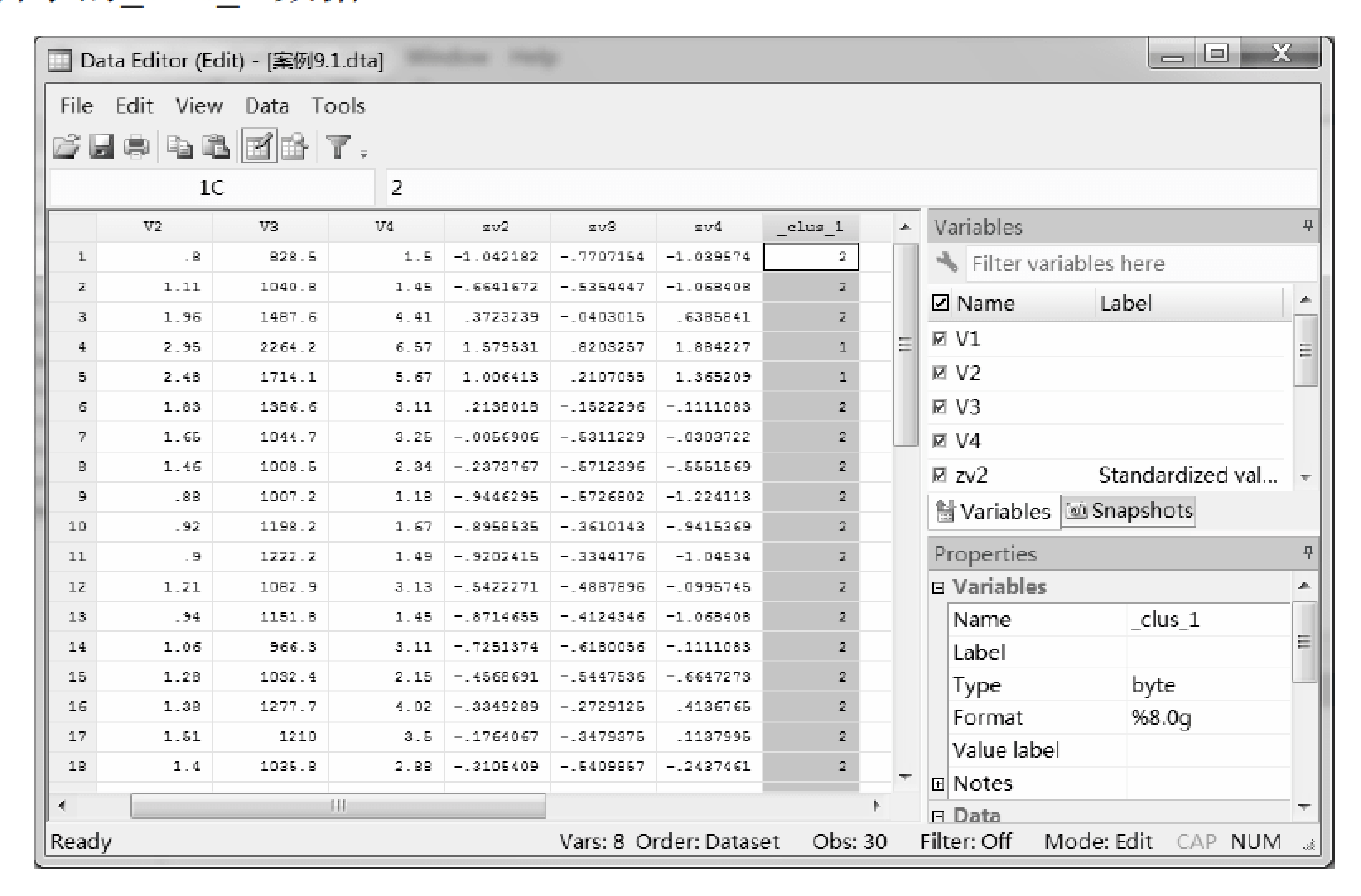
（1）设定聚类数为 2

图 9.4 展示的是设定聚类数为 2，然后使用“K 个平均数的聚类分析”方法进行分析的结果。在输入第 5 条 Stata 命令并且分别按键盘上的回车键进行确认后，我们可以看到系统产生了一个新的变量，即聚类变量_clus_1（cluster name: _clus_1）。

```
. cluster kmeans zv2 zv3 zv4,k(2)
cluster name: _clus_1
```

图 9.4　设定聚类数为 2 的“K 个平均数的聚类分析”方法进行分析的结果

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 9.5 所示的_clus_1 数据。



	V2	V3	V4	zv2	zv3	zv4	_clus_1
1	.8	828.5	1.5	-1.042182	-.7707154	-1.039574	2
2	1.11	1040.8	1.45	-.6641672	-.5354447	-1.068408	2
3	1.96	1487.6	4.41	.3723239	-.0403015	.6385841	2
4	2.95	2264.2	6.57	1.579531	.8203257	1.884227	1
5	2.48	1714.1	5.67	1.006413	.2107055	1.365209	1
6	1.83	1386.6	3.11	.2138018	-.1522296	-.1111083	2
7	1.65	1044.7	3.25	-.0056906	-.5311229	-.0303722	2
8	1.46	1008.5	2.34	-.2373767	-.5712396	-.5551569	2
9	.88	1007.2	1.18	-.9446295	-.5726802	-1.224113	2
10	.92	1198.2	1.67	-.8958535	-.3610143	-.9415369	2
11	.9	1222.2	1.49	-.9202415	-.3344176	-1.04534	2
12	1.21	1082.9	3.13	-.5422271	-.4887896	-.0995745	2
13	.94	1151.8	1.45	-.8714655	-.4124346	-1.068408	2
14	1.06	966.3	3.11	-.7251374	-.6180056	-.1111083	2
15	1.28	1032.4	2.15	-.4568691	-.5447536	-.6647273	2
16	1.38	1277.7	4.02	-.3349289	-.2729125	.4136765	2
17	1.51	1210	3.5	-.1764067	-.3479375	.1137995	2
18	1.4	1036.8	2.88	-.3105409	-.5409857	-.2437461	2

图 9.5　_clus_1 数据

在图 9.5 中，我们可以看到所有的观测样本被分为两类：其中，山西、内蒙古、甘肃、青海、宁夏被分到第 1 类，其他的省市被分到第 2 类。我们可以看到第 1 类的特征是单位地区生产总值煤消耗量、单位地区生产总值电消耗量以及单位工业增加值煤消耗量都相对较高。我们可以把第 1 类称为高能耗省市，把第 2 类称为低能耗省市。

（2）设定聚类数为 3

图 9.6 展示的是设定聚类数为 3，然后使用“K 个平均数的聚类分析”方法进行分析的结果。在输入第 6 条 Stata 命令并且分别按键盘上的回车键进行确认后，我们可以看到系统产生了一个新的变量，即聚类变量_clus_2（cluster name: _clus_2）。

```
. cluster kmeans zv2 zv3 zv4,k(3)
cluster name: _clus_2
```

图 9.6　设定聚类数为 3 的“K 个平均数的聚类分析”方法进行分析的结果

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 9.7 所示的_clus_2 数据。

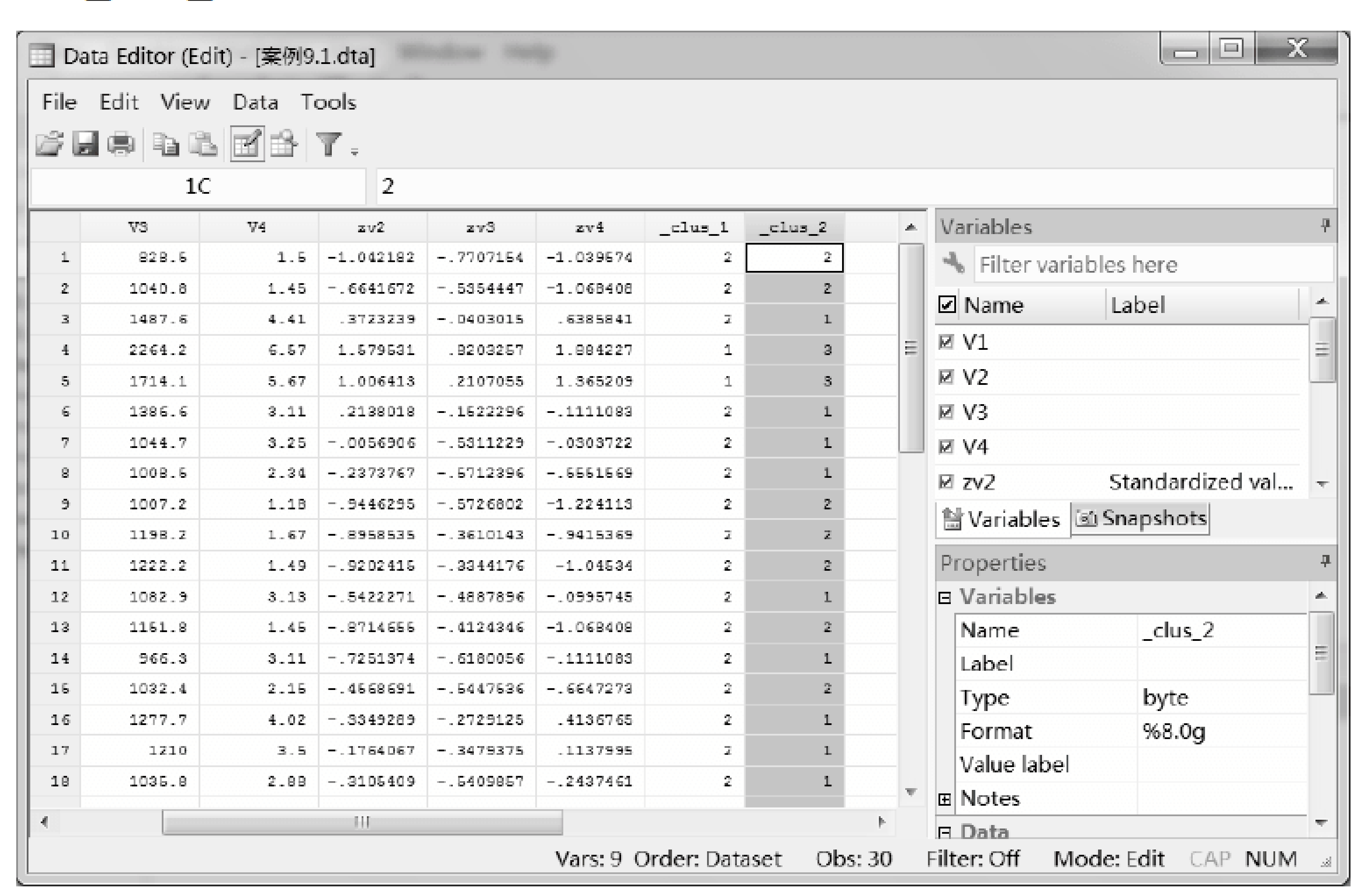


图 9.7　_clus_2 数据

在图 9.7 中，我们可以看到所有的观测样本被分为 3 类：其中，山西、内蒙古、贵州、甘肃、青海、宁夏被分到第 3 类；北京、天津、上海、江苏、浙江、福建、山东、广东被分到第 2 类；其他的省市被分到第 1 类。我们可以看到第 3 类的特征是单位地区生产总值煤消耗量、单位地区生产总值电消耗量以及单位工业增加值煤消耗量都较高，第 1 类的特征是单位地区生产总值煤消耗量、单位地区生产总值电消耗量以及单位工业增加值煤消耗量都处于中间，第 2 类的特征是单位地区生产总值煤消耗量、单位地区生产总值电消耗量以及单位工业增加值煤消耗量都较低。我们可以把第 3 类称为高能耗省市，把第 1 类称为中能耗省市，把第 2 类称为低能耗省市。

（3）设定聚类数为 4

图 9.8 展示的是设定聚类数为 4，然后使用“K 个平均数的聚类分析”方法进行分析的结果。在输入第 7 条 Stata 命令并且分别按键盘上的回车键进行确认后，我们可以看到系统产生了一个新的变量，即聚类变量_clus_3（cluster name: _clus_3）。

```
. cluster kmeans zv2 zv3 zv4,k(4)
cluster name: _clus_3
```

图 9.8　设定聚类数为 4 的“K 个平均数的聚类分析”方法进行分析的结果

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 9.9 所示的_clus_3 数据。

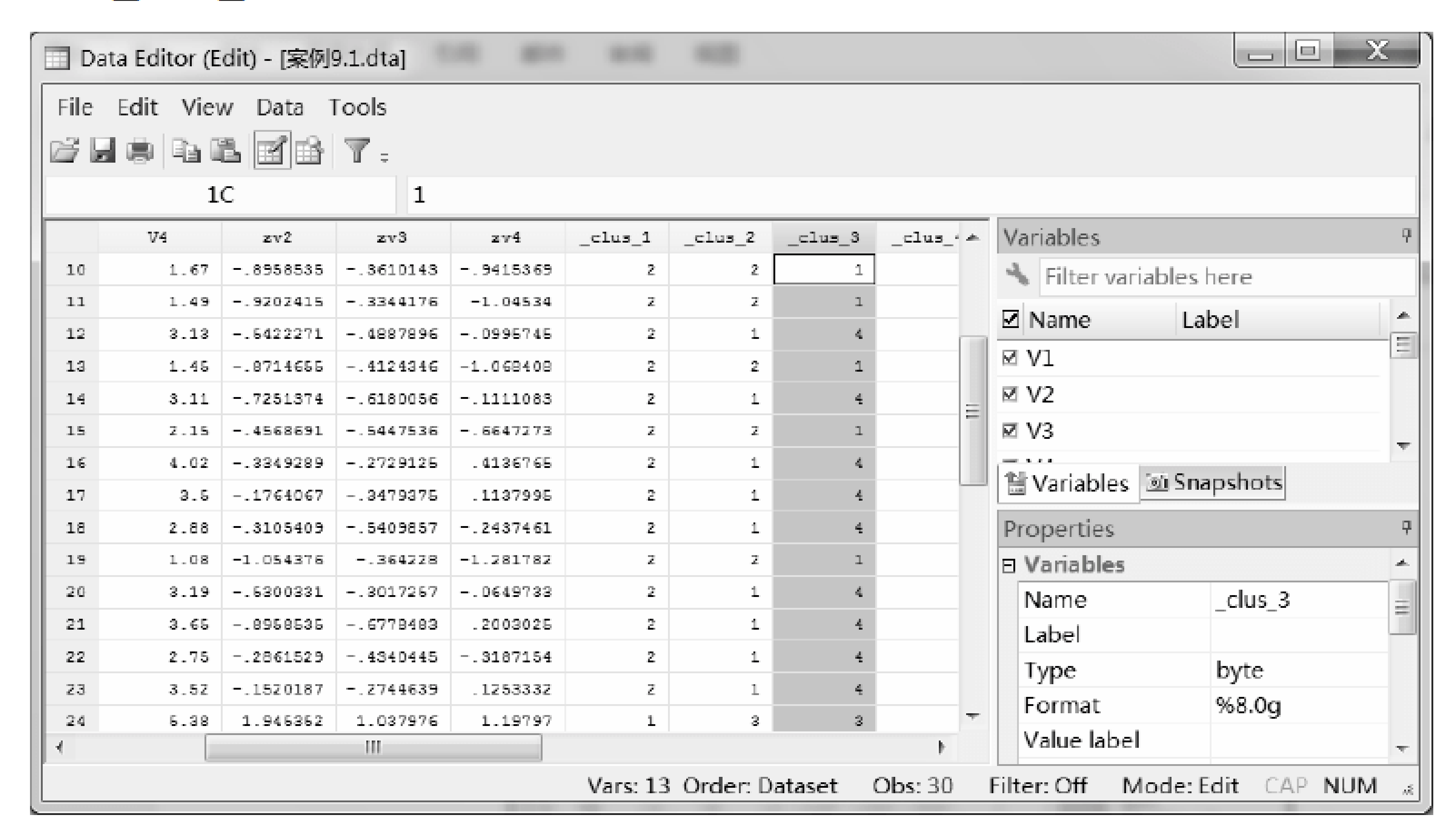


图 9.9　分析结果图

在图 9.9 中，可以看到所有的观测样本被分为 4 类：其中，北京、天津、上海、浙江、福建、江苏、广东、山东为第 1 类，宁夏、青海为第 2 类，甘肃、山西、贵州、内蒙古为第 3 类，其他省市为第 4 类。从图 9.9 中很难看出各个类别的特征，我们可以对数据进行排序操作，在主界面的“Command”文本框中输入操作命令：

```
sort _clus_3
```

并按键盘上的回车键进行确认，然后选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 9.10 所示的整理后的数据。

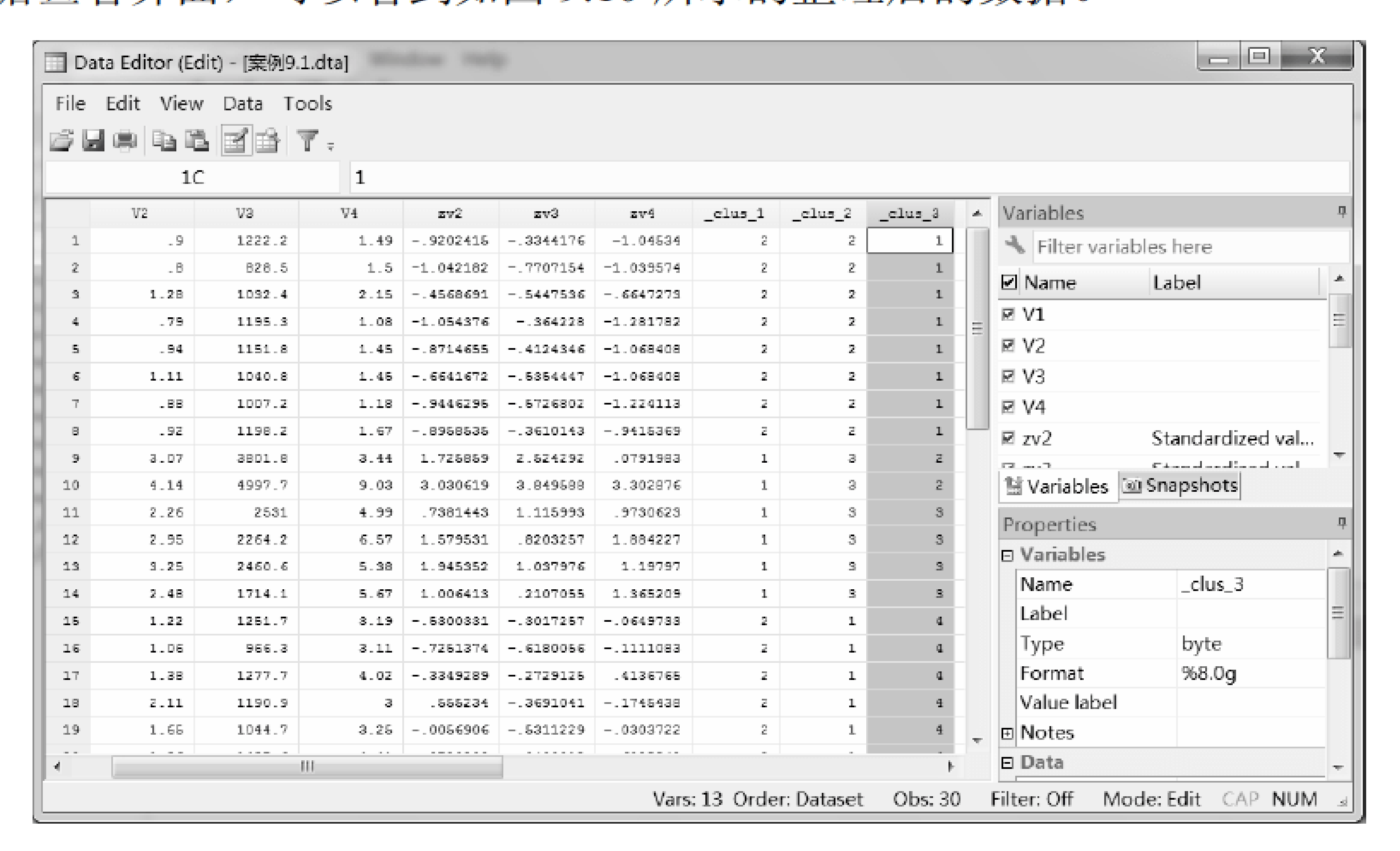


图 9.10　_clus_3 数据

从图 9.10 中可以看出，第 2 类的能耗应该是最高的，我们称为高能耗省市；然后是第 3 类，能耗较高，我们称为较高能耗省市；再后是第 4 类，能耗较低，我们称为较低能耗省市；第 1 类的能耗应该是最低的，我们称为低能耗省市。

在本节的开始我们也提到过，划分聚类分析的特点是需要事先制定拟分类的数量。究竟分成多少类是合理的，这是没有定论的。用户需要根据自己的研究、需要以及数据的实际特点加入自己的判断。在上面的分析中，我们尝试着把这 30 个样本分别分为 2、3、4 类进行了研究，我们可以看出把数据分成两类是过于粗糙的，而且两个类别所包含的样本数量差别也是比较大的，而把数据分成 3 类或者 4 类都是比较合适的。读者可以再把数据分成 5 类、6 类或者其他数量的类别进行研究，观察分类情况，取出自己认为是最优的分类。

3．K 个中位数的聚类分析

（1）设定聚类数为 2

图 9.11 展示的是设定聚类数为 2，然后使用“K 个中位数的聚类分析”方法进行分析的结果。在输入第 8 条 Stata 命令并且分别按键盘上的回车键进行确认后，我们可以看到系统产生了一个新的变量，即聚类变量_clus_4（cluster name: _clus_4）。

```
. cluster kmedians zv2 zv3 zv4,k(2)
cluster name: _clus_4
```

图 9.11　设定聚类数为 2 的“K 个中位数的聚类分析”方法进行分析的结果

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 9.12 所示的_clus_4 数据。

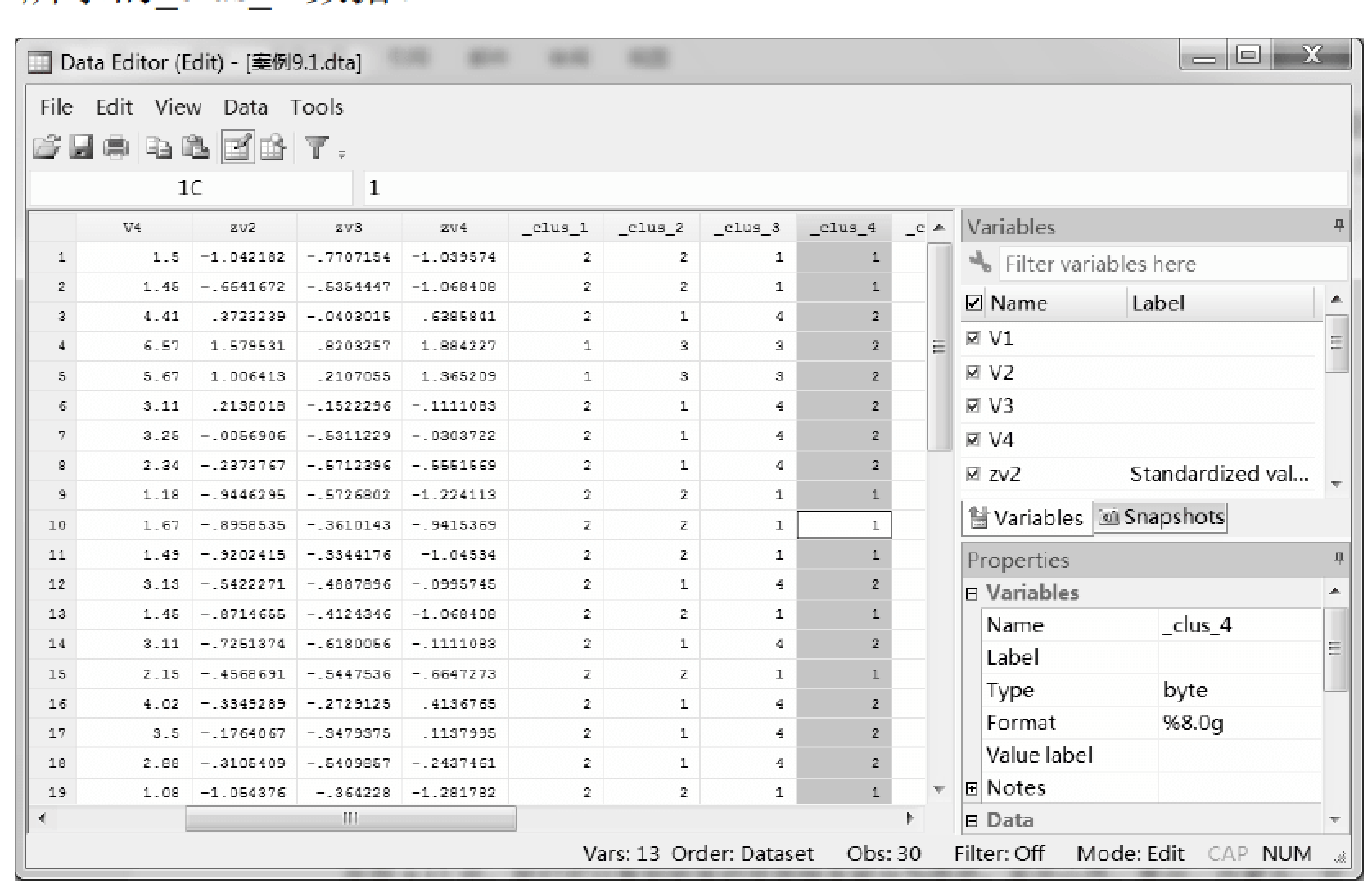


图 9.12　_clus_4 数据

在图 9.12 中，我们可以看到所有的观测样本被分为两类：其中，北京、天津、上海、江苏、浙江、广东、山东、福建被分到第 1 类，其他的省市被分到第 2 类。我们可以看到第 2 类的特征是单位地区生产总值煤消耗量、单位地区生产总值电消耗量以及单位工业增加值煤消耗量都相对非常高。我们可以把第 2 类称为高能耗省市，把第 1 类称为低能耗省市。

（2）设定聚类数为 3

图 9.13 展示的是设定聚类数为 3，然后使用“K 个中位数的聚类分析”方法进行分析的结

果。在输入第 9 条 Stata 命令并且分别按键盘上的回车键进行确认后，我们可以看到系统产生了一个新的变量，即聚类变量_clus_5（cluster name: _clus_5）。

```
. cluster kmedians zv2 zv3 zv4,k(3)
cluster name:  _clus_5
```

图 9.13　设定聚类数为 3 的“K 个中位数的聚类分析”方法进行分析的结果

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 9.14 所示的_clus_5 数据。

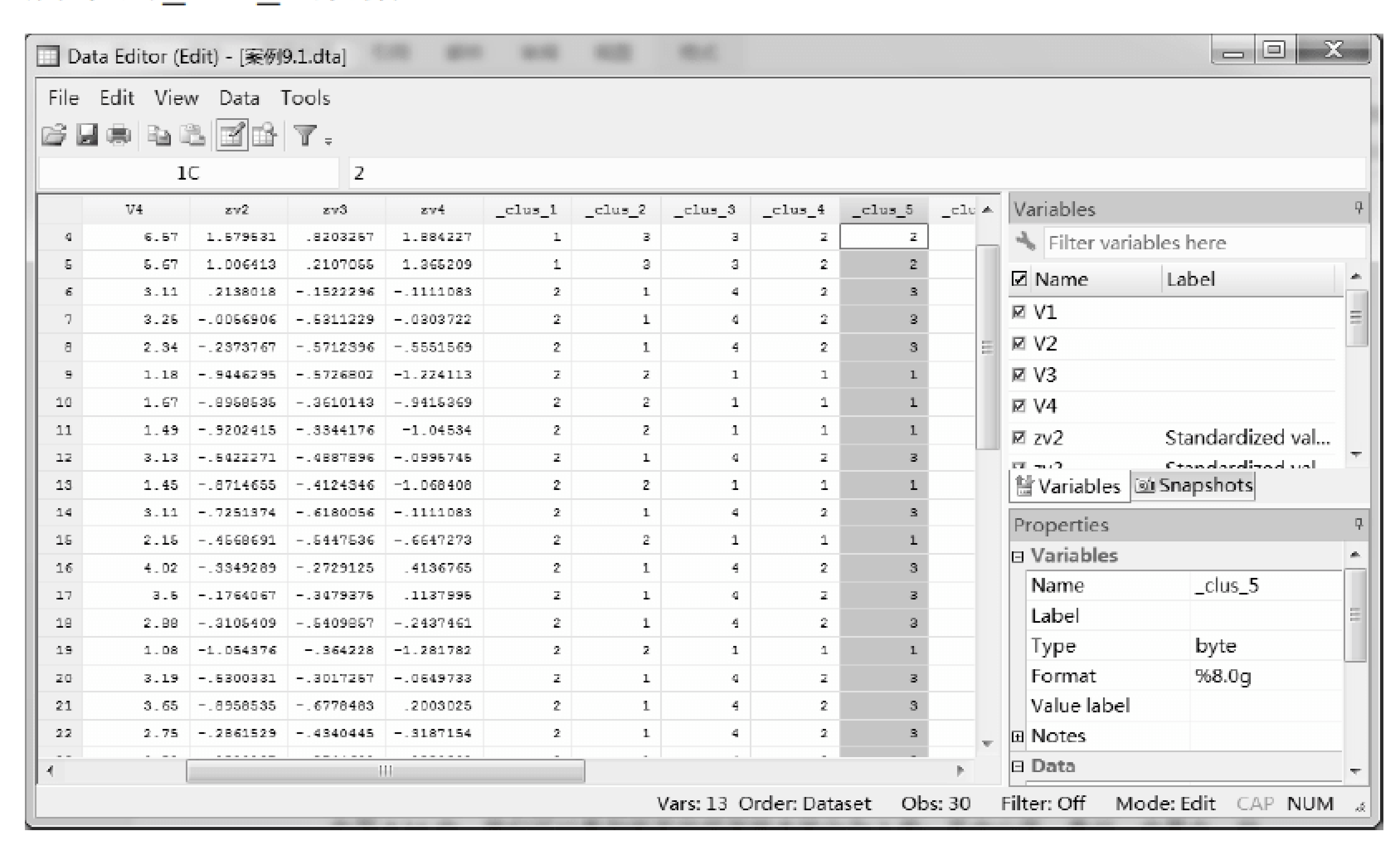


图 9.14　_clus_5 数据

在图 9.14 中，我们可以看到所有的观测样本被分为 3 类：其中，浙江、上海、福建、山东、北京、天津、广东、江苏被分到第 1 类，山西、贵州、内蒙古、甘肃、青海、宁夏被分到第 2 类，其他的省市被分到第 3 类。我们可以看到第 2 类的特征是单位地区生产总值煤消耗量、单位地区生产总值电消耗量以及单位工业增加值煤消耗量都较高，第 3 类的特征是单位地区生产总值煤消耗量、单位地区生产总值电消耗量以及单位工业增加值煤消耗量都处于中间，第 1 类的特征是单位地区生产总值煤消耗量、单位地区生产总值电消耗量以及单位工业增加值煤消耗量都较低。我们可以把第 2 类称为高能耗省市，把第 3 类称为中能耗省市，把第 1 类称为低能耗省市。

（3）设定聚类数为 4

图 9.15 展示的是设定聚类数为 4，然后使用“K 个中位数的聚类分析”方法进行分析的结果。在输入第 10 条 Stata 命令并且分别按键盘上的回车键进行确认后，我们可以看到系统产生了一个新的变量，即聚类变量_clus_6（cluster name: _clus_6）。

```
. cluster kmedians zv2 zv3 zv4,k(4)
cluster name:  _clus_6
```

图 9.15　设定聚类数为 4 的“K 个中位数的聚类分析”方法进行分析的结果

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 9.16 所示的_clus_6 数据。

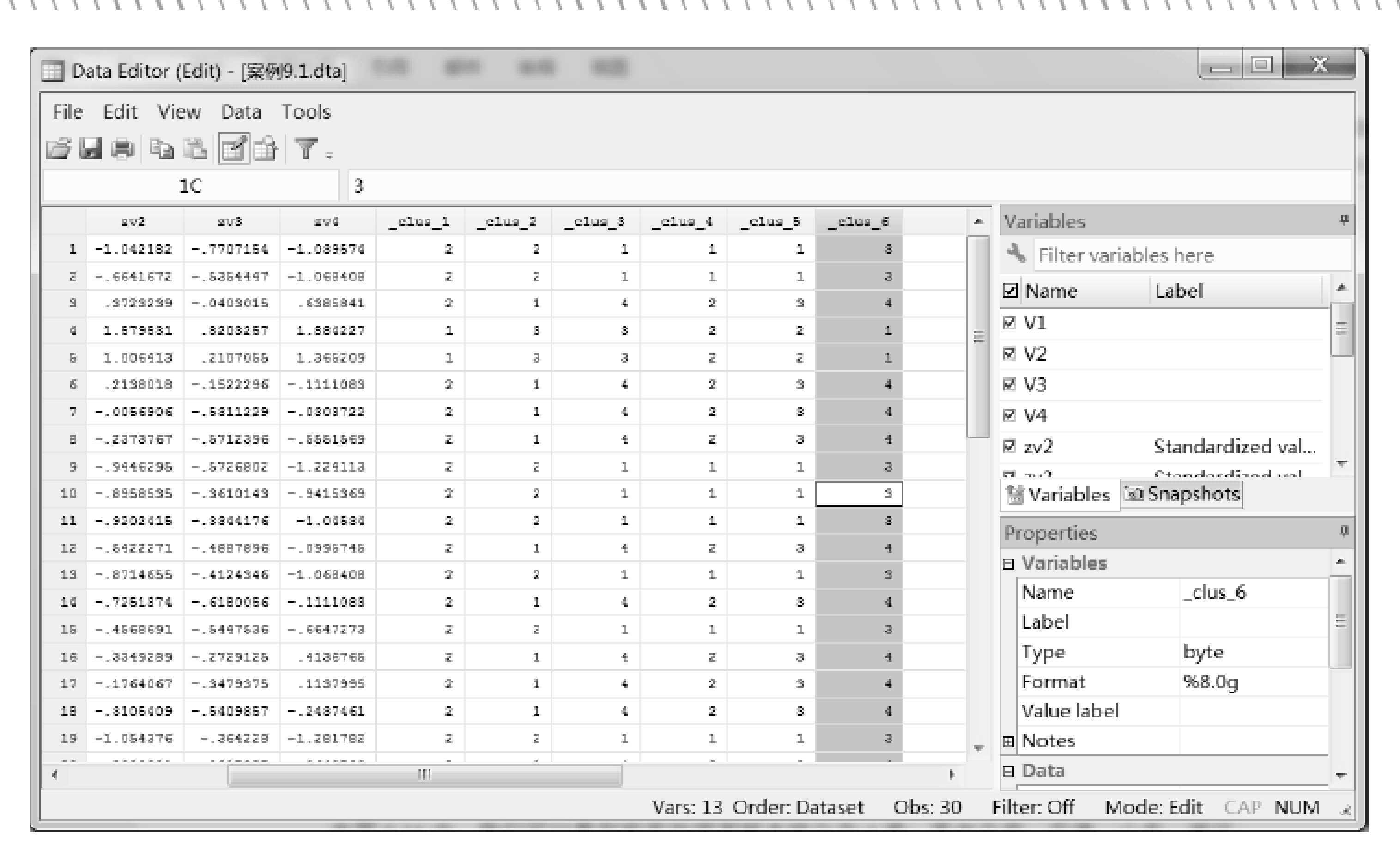


图 9.16 _clus_6 数据

在图 9.16 中，我们可以看到所有的观测样本被分为 4 类：其中，甘肃、青海、山西、贵州、内蒙古为第 1 类，宁夏为第 2 类，北京、天津、山东、浙江、上海、福建、江苏、广东为第 3 类，河北、新疆、辽宁、云南为第 3 类，其他省市为第 4 类。从图 9.16 中很难看出各个类别的特征，我们可以对数据进行排序操作，在主界面的“Command”文本框中输入操作命令：

```
sort _clus_6
```

并按键盘上的回车键进行确认，然后选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 9.17 所示的整理后的数据。

从图 9.17 中可以看出，第 2 类的能耗应该是最高的，我们称为高能耗省市；然后是第 1 类，能耗较高，我们称为较高能耗省市；再后是第 4 类，能耗较低，我们称为较低能耗省市；第 3 类的能耗应该是最低的，我们称为低能耗省市。

可以发现两种划分聚类分析方法得出的结论并不是完全一致的。关于两种方法孰优孰劣的问题，目前还没有定论，只是 K 个平均数的聚类分析方法应用更多一些。在实践过程中，用户可以根据研究的需要和自己的偏好进行选择，当然也可以同时将两种方法结合在一起进行综合判断。

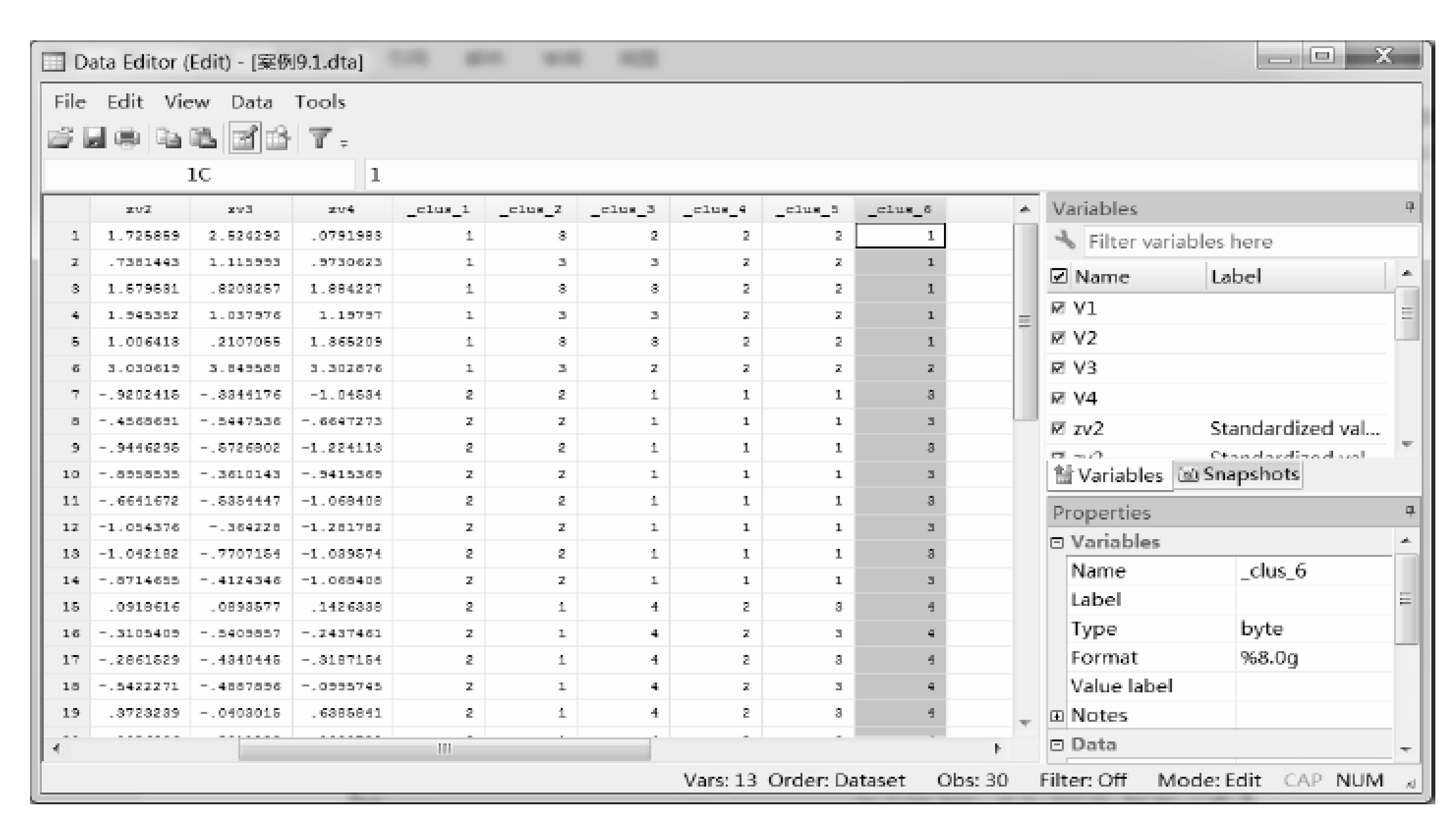


图 9.17 排序后_clus_6 数据

9.1.5　案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

1．延伸 1：采用其他相异性指标

在上面的实例中，聚类分析使用的相异性指标是系统的默认选项，也就是欧氏距离（Euclidean Distance）。除此之外，还有其他基于连续变量观测量的相异性指标可以使用，包括欧氏距离的平方（Squared Euclidean Distance）、绝对值距离（Absolute-Value Distance）、最大值距离（Maximum-Value Distance）、相关系数相似性度量（Correlation Coefficient Similarity Measure）等。例如，设定聚类数为 2，然后使用“K 个平均数的聚类分析”方法，采用欧氏距离的平方这一相异性指标，操作命令应该相应地修改为：

```
cluster kmeans zv2 zv3 zv4,k(2) measure(L2squared)
```

在命令窗口中输入命令并按回车键进行确认，结果如图 9.18~图 9.19 所示。

可以看到系统产生了一个新的变量，即聚类变量_clus_1（cluster name: _clus_1）。

```
. cluster kmeans zv2 zv3 zv4,k(2) measure(L2squared)
cluster name:  _clus_1
```

图 9.18　延伸 1 分析结果图

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 9.19 所示的_clus_1 数据。

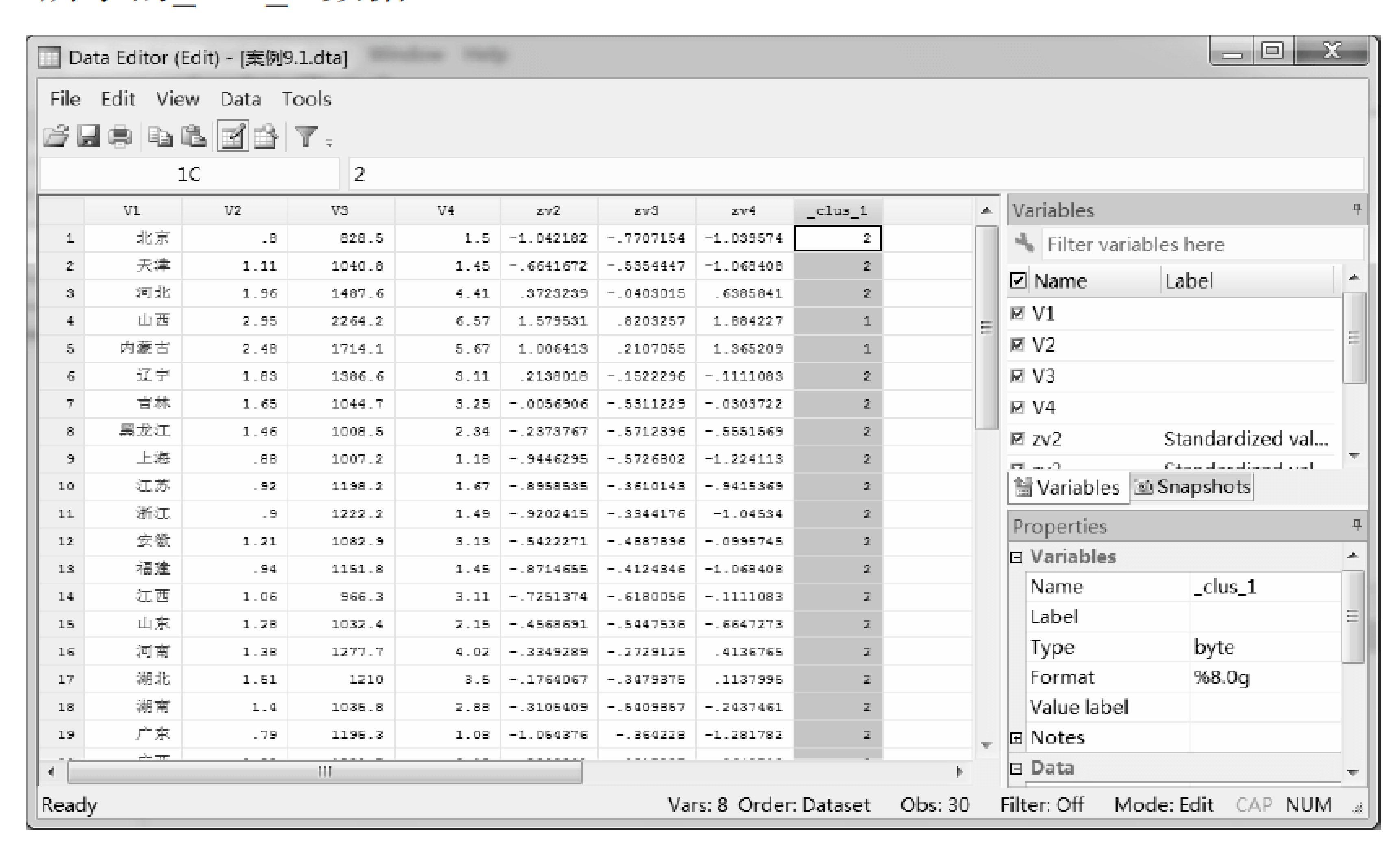


图 9.19　延伸 1 的_clus_1 数据

结果的解读方式与前面类似，限于篇幅，这里不再赘述。可以发现这两种测量方法下的聚类分析结果差别很大。基于连续变量观测量的相异性指标与对应的 Stata 14.0 命令如表 9.2 所示。

表 9.2　基于连续变量观测量的相异性指标与对应的 Stata 命令

基于连续变量观测量的相异性指标	对应的Stata命令
欧氏距离（Euclidean Distance）	L2
欧氏距离的平方（Squared Euclidean Distance）	L2squared
绝对值距离（Absolute-Value Distance）	L1
最大值距离（Maximum-Value Distance）	Linfinity
相关系数相似性度量（Correlation Coefficient Similarity Measure）	correlation

2．延伸 2：设置聚类变量的名称

在上面的实例中，聚类分析产生的聚类变量是系统默认生成的，例如_clus_1。事实上，我们可以个性化地设置聚类变量的名称。

例如，设定聚类数为 3，然后使用“K 个平均数的聚类分析”方法，采用绝对值距离的相异性指标，把产生的聚类变量取名为 abs，那么操作命令应该相应地修改为：

```
cluster kmeans zv2 zv3 zv4,k(3) measure(L1) name(abs)
```

在命令窗口输入命令并按回车键进行确认，然后选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 9.20 所示的 abs 数据。

图 9.20　延伸 2 分析结果图

结果的解读方式与前面类似，限于篇幅，这里不再赘述。

3．延伸 3：设置观测样本为初始聚类中心

可以根据拟聚类数，设置前几个观测样本为初始聚类中心进行聚类。

例如，设定聚类数为 3，然后使用“K 个平均数的聚类分析”方法，采用绝对值距离的相异性指标，把产生的聚类变量取名为 abcd，设置前几个观测样本为初始聚类中心进行聚类。那么操作命令应该相应地修改为：

```
cluster kmeans zv2 zv3 zv4,k(3)  measure(L1) name(abcd) start(firstk)
```

在命令窗口输入命令并按回车键进行确认，然后选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 9.21 所示的 abcd 数据。

结果的解读方式与前面类似，限于篇幅，这里不再赘述。

图 9.21 延伸 3 分析结果图

4．延伸 4：排除作为初始聚类中心的观测样本

在上面的实例中，我们可以根据拟聚类数，设置前几个观测样本为初始聚类中心进行聚类，但是在聚类分析时需要把作为初始聚类中心的观测样本排除。

例如，设定聚类数为 3，然后使用“K 个平均数的聚类分析”方法，采用绝对值距离的相异性指标，把产生的聚类变量取名为 abcde，设置前几个观测样本为初始聚类中心进行聚类，但是在聚类分析时需要把作为初始聚类中心的观测样本排除，那么操作命令应该相应地修改为：

```
cluster kmeans zv2 zv3 zv4,k(3)  measure(L1) name(abcde) start(firstk, exclude)
```

在命令窗口输入命令并按回车键进行确认，然后选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 9.22 所示的 abcde 数据。

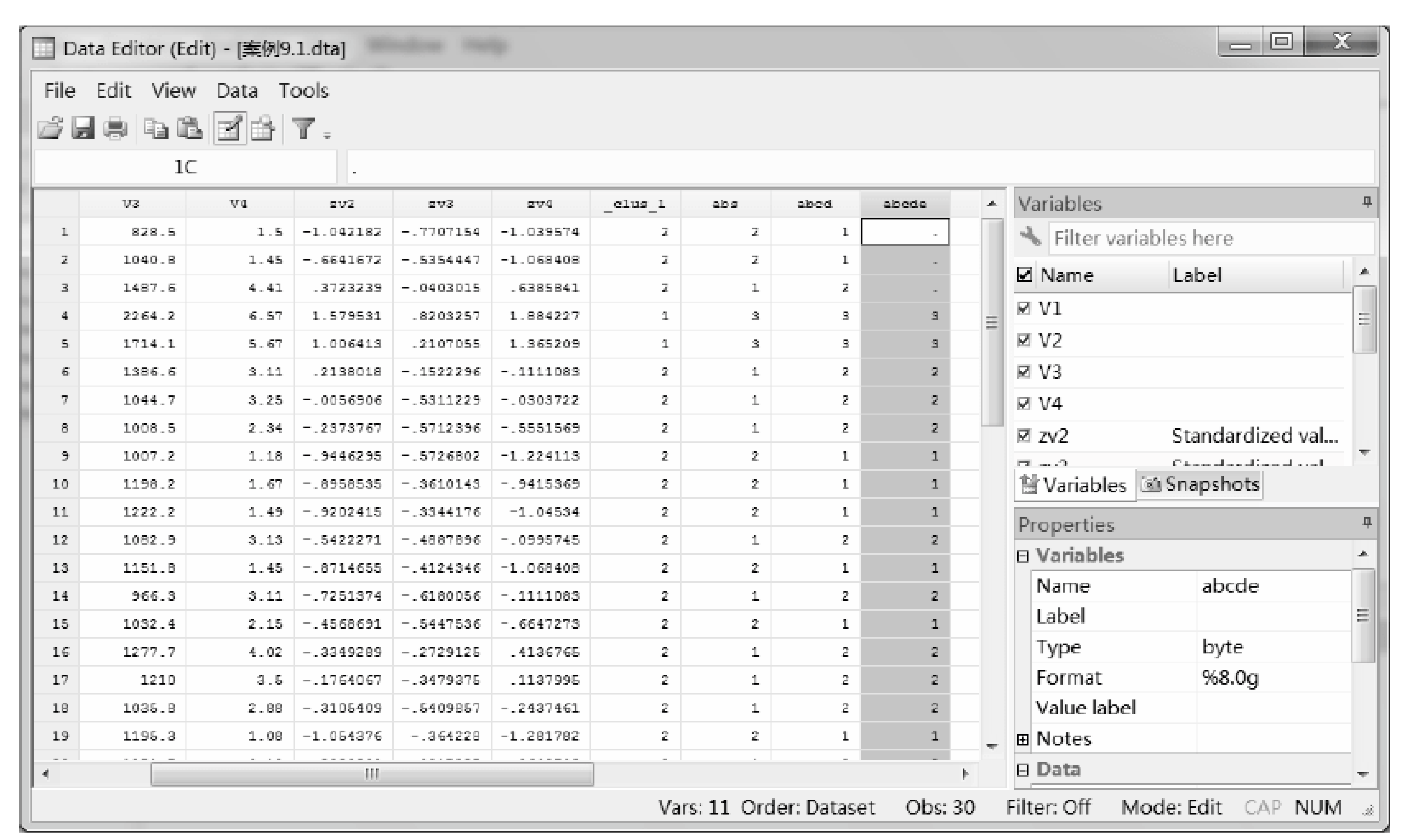


图 9.22 延伸 4 分析结果图

结果的解读方式与前面类似，限于篇幅，这里不再赘述。

9.2 实例二——层次聚类分析

9.2.1 层次聚类分析的功能与意义

层次聚类分析方法（Hierarchical）与划分聚类分析方法的原理不同，它的基本思想是根据一定的标准使得最相近的样本聚合到一起，然后逐步放松标准使得次相近的样本聚合到一起，最终实现完全聚类，即把所有的观测样本汇集到一个组的一种聚类方法。与划分聚类分析方法相比，层次聚类分析方法的计算过程更为复杂，计算速度相对较慢，但是它不要求事先指定需要分类的数量，这一点是符合聚类分析探索性的本质特点的，所以这种聚类分析方法应用也非常广泛。

9.2.2 相关数据来源

	下载资源:\video\chap09\···
	下载资源:\sample\chap09\案例9.2.dta

【例 9.2】党的十八大报告指出要千方百计增加居民收入，要提高居民收入在国民收入分配中的比重，要提高劳动报酬在初次分配中的比重。表 9.3 是我国 2005 年各地城镇居民平均每人全年家庭收入来源统计表。按照相关统计口径，各地城镇居民家庭收入来源分为工薪收入、经营净收入、财产性收入、转移性收入 4 个方面。试用层次聚类分析方法对全国各地区的收入来源结构进行分类，并进行简要论述分析。

表 9.3 2005 年各地区城镇居民每人全年家庭收入统计表（单位：元）

地区	工薪收入	经营净收入	财产性收入	转移性收入
北京	13 666.34	213.7	190.44	5 462.85
天津	8 174.64	665.53	148.15	4 574.99
河北	6 346.53	643.84	117.46	2 508.96
山西	7 103.45	350.96	136.38	1 947.77
…	…	…	…	…
甘肃	6 486.84	373.84	39.58	1 837.84
青海	5 613.79	513.41	62.08	2 577.4
宁夏	5 771.58	956.65	64.44	1 952.2
新疆	6 553.47	522.14	54.51	1 563.54

9.2.3 Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有 5 个变量，分别是地

区、工薪收入、经营净收入、财产性收入、转移性收入。我们把这些变量分别定义为 V1、V2、V3、V4、V5，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作我们在第 1 章中已有详细讲述。录入完成后数据如图 9.23 所示。

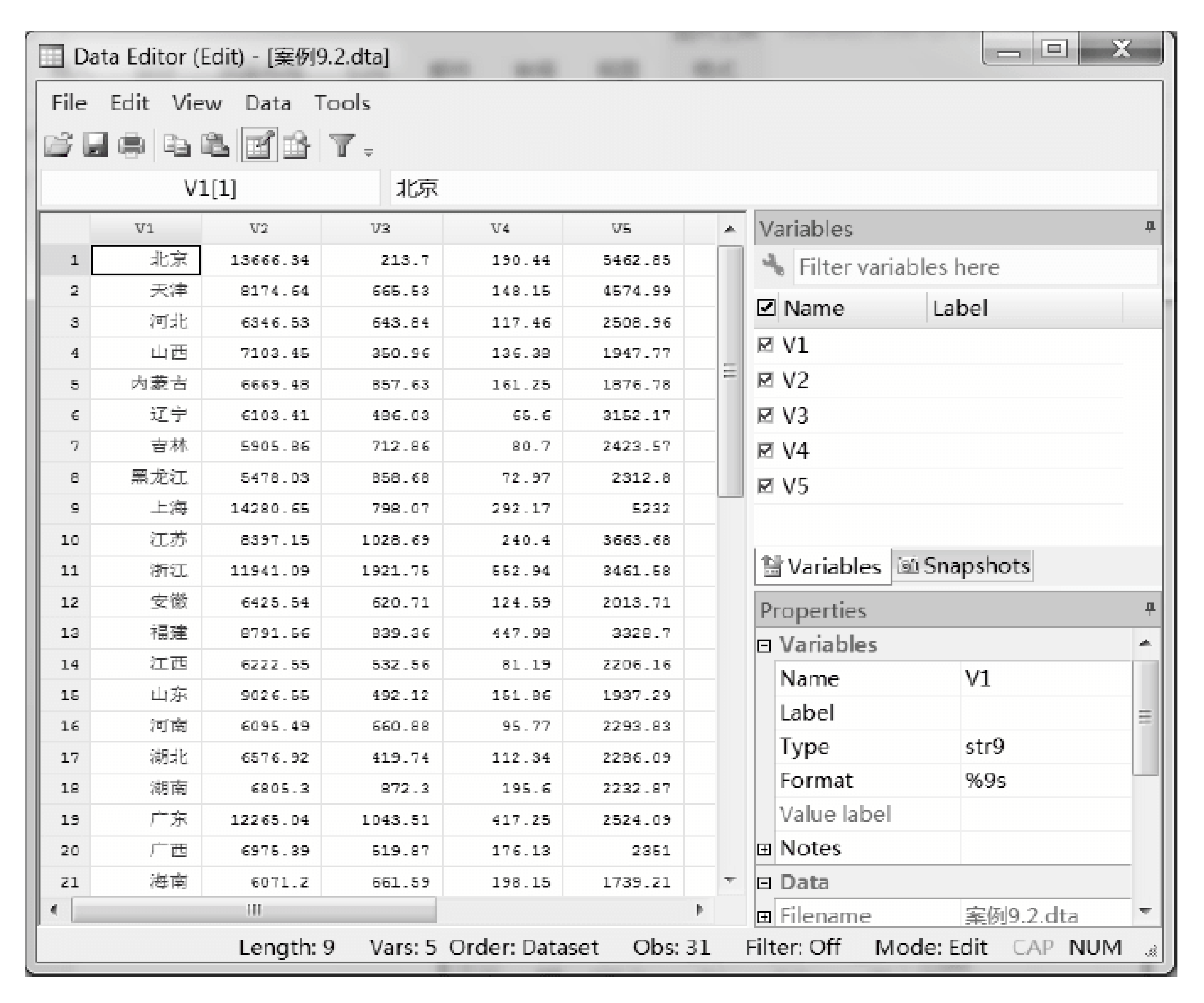


图 9.23　案例 9.2 数据

层次聚类分析方法（Hierarchical）有很多种，Stata 14.0 支持 7 种，包括最短联结法聚类分析（Single-Linkage Cluster Analysis）、最长联结法聚类分析（Complete-Linkage Cluster Analysis）、平均联结法聚类分析（Average-Linkage Cluster Analysis）、加权平均联结法聚类分析（Weighted-Average Linkage Cluster Analysis）、中位数联结法聚类分析（Median-Linkage Cluster Analysis）、重心联结法聚类分析（Centroid-Linkage Cluster Analysis）、Ward 联结法聚类分析（Ward's Linkage Cluster Analysis）等。我们先做一下数据保存，然后开始展开分析。

1．最短联结法聚类分析

操作步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中分别输入如下命令并按键盘上的回车键进行确认。

- egen zv2=std(V2)：本命令旨在对 V2 变量进行标准化处理。
- egen zv3=std(V3)：本命令旨在对 V3 变量进行标准化处理。
- egen zv4=std(V4)：本命令旨在对 V4 变量进行标准化处理。
- egen zv5=std(V5)：本命令旨在对 V5 变量进行标准化处理。
- summ　zv2 zv3 zv4 zv5：本命令旨在对 zv2、zv3、zv4、zv5 变量进行描述性统计分析。
- cluster singlelinkage　zv2 zv3 zv4 zv5：本命令旨在使用最短联结法对 zv2、zv3、zv4、zv5 变量进行层次聚类分析。

- cluster dendrogram：本命令旨在产生聚类分析树状图来描述层次聚类分析的结果。

03 设置完毕后，等待输出结果。

2．最长联结法聚类分析

操作步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。
02 在主界面的“Command”文本框中分别输入如下命令并按键盘上的回车键进行确认。

- egen zv2=std(V2)：本命令旨在对 V2 变量进行标准化处理。
- egen zv3=std(V3)：本命令旨在对 V3 变量进行标准化处理。
- egen zv4=std(V4)：本命令旨在对 V4 变量进行标准化处理。
- egen zv5=std(V5)：本命令旨在对 V5 变量进行标准化处理。
- summ zv2 zv3 zv4 zv5：本命令旨在对 zv2、zv3、zv4、zv5 变量进行描述性统计分析。
- cluster completelinkage zv2 zv3 zv4 zv5：本命令旨在使用最长联结法对 zv2、zv3、zv4、zv5 变量进行层次聚类分析。
- cluster dendrogram：本命令旨在产生聚类分析树状图来描述层次聚类分析的结果。

03 设置完毕后，等待输出结果。

3．平均联结法聚类分析

操作步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。
02 在主界面的“Command”文本框中分别输入如下命令并按键盘上的回车键进行确认。

- egen zv2=std(V2)：本命令旨在对 V2 变量进行标准化处理。
- egen zv3=std(V3)：本命令旨在对 V3 变量进行标准化处理。
- egen zv4=std(V4)：本命令旨在对 V4 变量进行标准化处理。
- egen zv5=std(V5)：本命令旨在对 V5 变量进行标准化处理。
- summ zv2 zv3 zv4 zv5：本命令旨在对 zv2、zv3、zv4、zv5 变量进行描述性统计分析。
- cluster averagelinkage zv2 zv3 zv4 zv5：本命令旨在使用平均联结法对 zv2、zv3、zv4、zv5 变量进行层次聚类分析。
- cluster dendrogram：本命令旨在产生聚类分析树状图来描述层次聚类分析的结果。

03 设置完毕后，等待输出结果。

4．加权平均联结法聚类分析

操作步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中分别输入如下命令并按键盘上的回车键进行确认。

- egen zv2=std(V2)：本命令旨在对 V2 变量进行标准化处理。
- egen zv3=std(V3)：本命令旨在对 V3 变量进行标准化处理。
- egen zv4=std(V4)：本命令旨在对 V4 变量进行标准化处理。
- egen zv5=std(V5)：本命令旨在对 V5 变量进行标准化处理。
- summ　zv2 zv3 zv4 zv5：本命令旨在对 zv2、zv3、zv4、zv5 变量进行描述性统计分析。
- cluster waveragelinkage　zv2 zv3 zv4 zv5：本命令旨在使用加权平均联结法对 zv2、zv3、zv4、zv5 变量进行层次聚类分析。
- cluster dendrogram：本命令旨在产生聚类分析树状图来描述层次聚类分析的结果。

03 设置完毕后，等待输出结果。

5．中位数联结法聚类分析

操作步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。
02 在主界面的“Command”文本框中分别输入如下命令并按键盘上的回车键进行确认。

- egen zv2=std(V2)：本命令旨在对 V2 变量进行标准化处理。
- egen zv3=std(V3)：本命令旨在对 V3 变量进行标准化处理。
- egen zv4=std(V4)：本命令旨在对 V4 变量进行标准化处理。
- egen zv5=std(V5)：本命令旨在对 V5 变量进行标准化处理。
- summ　zv2 zv3 zv4 zv5：本命令旨在对 zv2、zv3、zv4、zv5 变量进行描述性统计分析。
- cluster medianlinkage zv2 zv3 zv4 zv5：本命令旨在使用中位数联结法对 zv2、zv3、zv4、zv5 变量进行层次聚类分析。
- cluster dendrogram：本命令旨在产生聚类分析树状图来描述层次聚类分析的结果。

03 设置完毕后，等待输出结果。

6．重心联结法聚类分析

操作步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。
02 在主界面的“Command”文本框中分别输入如下命令并按键盘上的回车键进行确认。

- egen zv2=std(V2)：本命令旨在对 V2 变量进行标准化处理。
- egen zv3=std(V3)：本命令旨在对 V3 变量进行标准化处理。
- egen zv4=std(V4)：本命令旨在对 V4 变量进行标准化处理。
- egen zv5=std(V5)：本命令旨在对 V5 变量进行标准化处理。
- summ　zv2 zv3 zv4 zv5：本命令旨在对 zv2、zv3、zv4、zv5 变量进行描述性统计分析。

- cluster centroidlinkage zv2 zv3 zv4 zv5：本命令旨在使用重心联结法对 zv2、zv3、zv4、zv5 变量进行层次聚类分析。

03 设置完毕后，等待输出结果。

7．Ward 联结法聚类分析

操作步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。
02 在主界面的"Command"文本框中分别输入如下命令并按键盘上的回车键进行确认。

- egen zv2=std(V2)：本命令旨在对 V2 变量进行标准化处理。
- egen zv3=std(V3)：本命令旨在对 V3 变量进行标准化处理。
- egen zv4=std(V4)：本命令旨在对 V4 变量进行标准化处理。
- egen zv5=std(V5)：本命令旨在对 V5 变量进行标准化处理。
- summ zv2 zv3 zv4 zv5：本命令旨在对 zv2、zv3、zv4、zv5 变量进行描述性统计分析。
- cluster wardslinkage zv2 zv3 zv4 zv5：本命令旨在使用 Ward 联结法对 zv2、zv3、zv4、zv5 变量进行层次聚类分析。
- cluster dendrogram：本命令旨在产生聚类分析树状图来描述层次聚类分析的结果。

03 设置完毕后，等待输出结果。

9.2.4 结果分析

在 Stata 14.0 主界面的结果窗口我们可以看到如图 9.24~图 9.45 所示的分析结果。

1．最短联结法聚类分析（Single-Linkage Cluster Analysis）

在分析过程中前 4 条 Stata 命令旨在对数据进行标准化处理，选择的标准化处理方式是使变量的平均数为 0 而且标准差为 1。之所以这样做是因为我们进行聚类分析的变量都是以不可比的单位进行的测度，它们具有极为不同的方差，我们对数据进行标准化处理可以避免使结果受到具有最大方差变量的影响。在输入前 4 条 Stata 命令并且分别按键盘上的回车键进行确认后，选择"Data" | "Data Editor" | "Data Editor(Browse)"命令，进入数据查看界面，可以看到如图 9.24 所示的变换后的数据。

	v1	v2	v3	v4	v5	zv2	zv3	zv4	zv5
1	北京	13666.34	213.7	190.44	5462.85	2.491832	-1.300932	.1381473	2.74791
2	天津	8174.64	665.53	148.15	4574.99	.2327827	.0430095	-.1851475	1.904427
3	河北	6346.53	643.84	117.46	2508.96	-.5192236	-.0215061	-.4197636	-.0583391
4	山西	7103.45	350.96	136.38	1947.77	-.2078589	-.8926602	-.2751256	-.5914798
5	内蒙古	6669.48	857.63	161.25	1876.78	-.3863757	.6143994	-.0850018	-.6589216
6	辽宁	6103.41	486.03	65.6	3152.17	-.6192325	-.4909025	-.8162181	.552722
7	吉林	5905.86	712.86	80.7	2423.57	-.7004963	.1837896	-.7007831	-.1394611
8	黑龙江	5478.03	858.68	72.97	2312.8	-.8764872	.6175225	-.7598766	-.2446946
9	上海	14280.65	798.07	292.17	5232	2.744534	.4372417	.9158435	2.528599
10	江苏	8397.15	1028.69	240.4	3663.68	.3243138	1.123207	.5200769	1.038666
11	浙江	11941.09	1921.75	552.94	3461.58	1.782139	3.779561	2.909354	.8466673
12	安徽	6425.54	620.71	124.59	2013.71	-.4867221	-.0903049	-.3652568	-.5288357
13	福建	8791.56	839.36	447.98	3328.7	.4865569	.5600563	2.106966	.7204288
14	江西	6222.55	532.56	81.19	2206.16	-.5702236	-.3525019	-.6970371	-.3460047
15	山东	9026.55	492.12	151.86	1937.29	.5832218	-.4727882	-.1567856	-.601436
16	河南	6095.49	660.88	95.77	2293.83	-.6224905	.0291783	-.5855773	-.2627164
17	湖北	6576.92	419.74	112.34	2286.09	-.4244509	-.6880782	-.4589045	-.2700696
18	湖南	6805.3	872.3	195.6	2232.87	-.3305052	.6580344	.177594	-.3206295
19	广东	12265.04	1043.51	417.25	2524.09	1.915398	1.167288	1.872044	-.0439652
20	广西	6975.39	519.87	176.13	2351	-.2605374	-.3902475	.0287515	-.208404
21	海南	6071.2	661.59	198.15	1739.21	-.6324823	.0312902	.1970879	-.7896157
22	重庆	7848.52	492.44	188.22	2549.97	.0986308	-.4718364	.121176	-.0193789
23	四川	5838.27	515.49	211.41	2438.41	-.7282998	-.4032756	.2984568	-.125363
24	贵州	5516.18	790.84	90.25	1987.8	-.8607938	.4157366	-.6277761	-.5534505
25	云南	6170.93	595.45	428.07	2800.2	-.5914577	-.1654393	1.954759	.2183442
26	西藏	10401.71	43.2	10.41	204	1.148904	-1.808074	-1.23813	-2.248093
27	陕西	6347.81	179.34	135.15	2239.96	-.5186969	-1.403134	-.2845287	-.313894
28	甘肃	6486.84	373.84	39.58	1837.84	-.461506	-.824605	-1.015133	-.6959153
29	青海	5613.79	513.41	62.08	2577.4	-.8206413	-.4094625	-.8431275	.0066801
30	宁夏	5771.58	956.65	64.44	1952.2	-.7557332	.9089285	-.825086	-.5872713
31	新疆	6553.47	522.14	54.51	1563.54	-.4340971	-.3834955	-.9009979	-.9565053

图 9.24　标准化变换后的数据

根据我们在前面章节中讲述的描述性统计分析方法，可以看到如图 9.25 所示的标准化变量的相应统计量。

```
. summ  zv2 zv3 zv4 zv5

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         zv2 |        31    2.40e-09           1  -.8764872   2.744534
         zv3 |        31    1.56e-09           1  -1.808074   3.779561
         zv4 |        31    1.08e-09           1   -1.23813   2.909354
         zv5 |        31   -5.86e-10           1  -2.248093    2.74791
```

图 9.25　标准化变量的相应统计量分析结果图

通过观察分析结果可以看出，有效观测样本共有 31 个。zv2 的平均值为 2.40e-09，标准差是 1，最小值是-0.8764872，最大值是 2.744534；zv3 的平均值为 1.56e-09，标准差是 1，最小值是-1.808074，最大值是 3.779561；zv4 的平均值为 1.08e-09，标准差是 1，最小值是-1.23813，最大值是 2.909354；zv5 的平均值为-5.86e-10，标准差是 1，最小值是-2.248093，最大值是 2.74791。

图 9.26 展示的是使用“最短联结法聚类分析”方法进行分析的结果。在输入第 6 条 Stata 命令并且分别按键盘上的回车键进行确认后，可以看到系统产生了一个新的变量，即聚类变量_clus_1（cluster name: _clus_1）。



```
. cluster singlelinkage  zv2 zv3 zv4 zv5
cluster name:  _clus_1
```

图 9.26　最短联结法聚类分析结果图

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 9.27 所示的_clus_1 数据。

	V1	zv2	zv3	zv4	zv5	_clus_1_id	_clus_1_ord	_clus_1_hgt	type1	type2
1	北京	2.491832	-1.300932	.1381473	2.74791	1	11	2.9514342	2	2
2	天津	.2327827	.0430095	-.1851475	1.904427	2	1	1.9333945	4	2
3	河北	-.5192236	-.0215061	-.4197636	-.0583391	3	9	2.7794725	4	2
4	山西	-.2078589	-.8926602	-.2751256	-.5914798	4	26	2.4461742	4	2
5	内蒙古	-.3863757	.6143994	-.0850018	-.6589216	5	19	1.7463687	4	2
6	辽宁	-.6192325	-.4909025	-.8162181	.552722	6	13	1.4013262	4	2
7	吉林	-.7004963	.1837896	-.7007831	-.1394611	7	25	1.7136991	4	2
8	黑龙江	-.8764872	.6175225	-.7598766	-.2446946	8	2	1.556306	4	2
9	上海	2.744534	.4372417	.9158435	2.528599	9	10	1.6155965	2	2
10	江苏	.3243138	1.123207	.5200769	1.038666	10	15	.80677284	4	2
11	浙江	1.782139	3.779561	2.909354	.8466673	11	5	.43407789	1	1
12	安徽	-.4867221	-.0903049	-.3652568	-.5288357	12	18	.70515268	4	2
13	福建	.4865569	.5600563	2.106966	.7204288	13	27	.65904896	4	2
14	江西	-.5702236	-.3525019	-.6970371	-.3460047	14	21	.64828178	4	2
15	山东	.5832218	-.4727882	-.1567856	-.601436	15	20	.42418311	4	2
16	河南	-.6224905	.0291783	-.5855773	-.2627164	16	22	.54645054	4	2
17	湖北	-.4244509	-.6880782	-.4589045	-.2700696	17	23	.59764691	4	2
18	湖南	-.3305052	.6580344	.177594	-.3206295	18	6	.58828898	4	2
19	广东	1.915398	1.167288	1.872044	-.0439652	19	4	.47523163	4	2
20	广西	-.2605374	-.3902475	.0287515	-.208404	20	14	.40959957	4	2
21	海南	-.6324823	.0312902	.1970879	-.7896157	21	12	.38996133	4	2
22	重庆	.0986308	-.4718364	.121176	-.0193789	22	3	.28722333	4	2
23	四川	-.7282998	-.4032756	.2984568	-.125363	23	7	.24177225	4	2
24	贵州	-.8607938	.4157366	-.6277761	-.5534505	24	16	.44309643	4	2
25	云南	-.5914577	-.1654393	1.954759	.2183442	25	17	.46008964	4	2
26	西藏	1.148904	-1.808074	-1.23813	-2.248093	26	29	.48338714	3	2
27	陕西	-.5186969	-1.403134	-.2845287	-.313894	27	8	.39210278	4	2
28	甘肃	-.461506	-.824605	-1.015133	-.6959153	28	24	.47022339	4	2
29	青海	-.8206413	-.4094625	-.8431275	.0066801	29	30	.65863648	4	2
30	宁夏	-.7557332	.9089285	-.825086	-.5872713	30	28	.52560714	4	2
31	新疆	-.4340971	-.3834955	-.9009979	-.9565053	31	31	.	4	2

图 9.27　_clus_1 数据

在图 9.27 中，可以看到层次聚类分析方法产生的聚类变量是与划分聚类分析方法不同的，它包括 3 个组成部分：_clus_1_id、_clus_1_ord、_clus_1_hgt。其中，_clus_1_id 表示的是系统对该观测样本的初始编号；_clus_1_ord 表示的是系统对该观测样本进行聚类分析处理后的编号；_clus_1_hgt 表示的是系统对该观测样本进行聚类计算后的值。

为了使聚类分析的结果可视化，我们需要绘制如图 9.28 所示的聚类分析树状图。在输入第 7 条 Stata 命令并且分别按键盘上的回车键进行确认后，可以看到系统产生了聚类分析树状图。

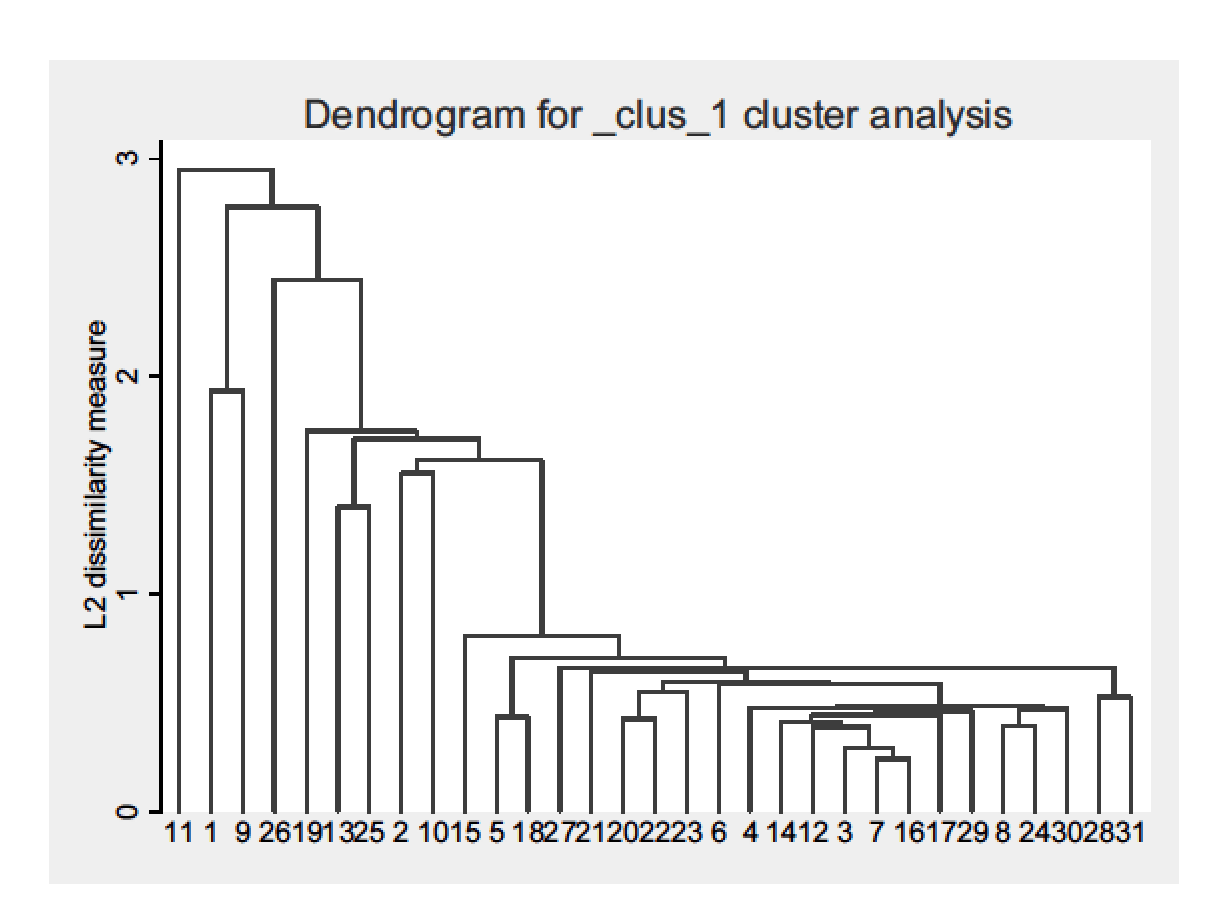


图 9.28　聚类分析树状图

观察图 9.28，可以直观地看到具体的聚类情况：7 号样本跟 16 号样本首先聚合在一起，进入数据查看界面查看_clus_1_id 变量， 7 号样本代表的是吉林，16 号样本代表的是河南。7 号样本与 16 号样本聚合后又与 3 号样本（河北）聚合，依次类推，最后 11 号样本（浙江）与所有样本聚合为一类。那么，到底分成了多少类呢？答案是不确定的，因为这取决于研究的需

要和实际的情况，需要用户加入自己的判断。例如，可分成两类，即 11 号样本（浙江）单独一类，其他的样本属于一类。

2．最长联结法聚类分析（Complete-Linkage Cluster Analysis）

在分析过程中前 4 条 Stata 命令旨在对数据进行标准化处理，选择的标准化处理方式是使变量的平均数为 0，而且标准差为 1。处理结果与最短联结法聚类分析是一致的，限于篇幅，这里不再赘述。

图 9.29 展示的是使用“最长联结法聚类分析”方法进行分析的结果。在输入第 6 条 Stata 命令并且分别按键盘上的回车键进行确认后，可以看到系统产生了一个新的变量，即聚类变量 _clus_1（cluster name: _clus_1）。

```
. cluster completelinkage zv2 zv3 zv4 zv5
cluster name: _clus_1
```

图 9.29 最长联结法聚类分析结果图

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 9.30 所示的_clus_1 数据。

	v2	v3	v4	v5	zv2	zv3	zv4	zv5	_clus_1_id	_clus_1_ord	_clus_1_hgt
1	13666.34	213.7	190.44	5462.85	2.491832	-1.300932	.1381473	2.74791	1	1	1.9333945
2	8174.64	665.53	148.15	4574.99	.2327827	.0430095	-.1851475	1.904427	2	9	5.9197855
3	6346.53	643.84	117.46	2508.96	-.5192236	-.0215061	-.4197636	-.0583391	3	2	1.556306
4	7103.45	350.96	136.38	1947.77	-.2078589	-.8926602	-.2751256	-.5914798	4	10	3.1268397
5	6669.48	857.63	161.25	1876.78	-.3863757	.6143994	-.0850018	-.6589216	5	3	.40069796
6	6103.41	486.03	65.6	3152.17	-.6192325	-.4909025	-.8162181	.552722	6	7	.24177225
7	5905.86	712.86	80.7	2423.57	-.7004963	.1837896	-.7007831	-.1394611	7	16	.62049788
8	5478.03	858.68	72.97	2312.8	-.8764872	.6175225	-.7598766	-.2446946	8	12	1.0798934
9	14280.65	798.07	292.17	5232	2.744534	.4372417	.9158435	2.528599	9	5	.43407789
10	8397.15	1028.69	240.4	3663.68	.3243138	1.123207	.5200769	1.038666	10	18	1.168632
11	11941.09	1921.75	552.94	3461.58	1.782139	3.779561	2.909354	.8466673	11	8	.39210278
12	6425.54	620.71	124.59	2013.71	-.4867221	-.0903049	-.3652568	-.5288357	12	24	.54254129
13	8791.56	839.36	447.98	3328.7	.4865569	.5600563	2.106966	.7204288	13	30	2.4018231
14	6222.55	532.56	81.19	2206.16	-.5702236	-.3525019	-.6970371	-.3460047	14	4	.65904896
15	9026.55	492.12	151.86	1937.29	.5832218	-.4727882	-.1567856	-.601436	15	27	1.3563932
16	6095.49	660.88	95.77	2293.83	-.6224905	.0291783	-.5855773	-.2627164	16	14	.44309643
17	6576.92	419.74	112.34	2286.09	-.4244509	-.6880782	-.4589045	-.2700696	17	17	.87149546
18	6805.3	872.3	195.6	2232.87	-.3305052	.6580344	.177594	-.3206295	18	28	.52560714
19	12265.04	1043.51	417.25	2524.09	1.915398	1.167288	1.872044	-.0439652	19	31	1.5266847
20	6975.39	519.87	176.13	2351	-.2605374	-.3902475	.0287515	-.208404	20	6	.58828898
21	6071.2	661.59	198.15	1739.21	-.6324823	.0312902	.1970879	-.7896157	21	29	1.792528
22	7848.52	492.44	188.22	2549.97	.0986308	-.4718364	.121176	-.0193789	22	15	.95270175
23	5838.27	515.49	211.41	2438.41	-.7282998	-.4032756	.2984568	-.125363	23	20	.42418311
24	5516.18	790.84	90.25	1987.8	-.8607938	.4157366	-.6277761	-.5534505	24	22	1.4692896
25	6170.93	595.45	428.07	2800.2	-.5914577	-.1654393	1.954759	.2183442	25	21	.80593782
26	10401.71	43.2	10.41	204	1.148904	-1.808074	-1.23813	-2.248093	26	23	4.2869733
27	6347.81	179.34	135.15	2239.96	-.5186969	-1.403134	-.2845287	-.313894	27	13	1.4013262
28	6486.84	373.84	39.58	1837.84	-.461506	-.824605	-1.015133	-.6959153	28	25	2.8523912
29	5613.79	513.41	62.08	2577.4	-.8206413	-.4094625	-.8431275	.0066801	29	19	5.1037758
30	5771.58	956.65	64.44	1952.2	-.7557332	.9089285	-.825086	-.5872713	30	26	7.6421076
31	6553.47	522.14	54.51	1563.54	-.4340971	-.3834955	-.9009979	-.9565053	31	11	.

图 9.30 _clus_1 数据

为了使聚类分析的结果可视化，我们需要绘制如图 9.31 所示的聚类分析树状图。在输入第 7 条 Stata 命令并且分别按键盘上的回车键进行确认后，可以看到系统产生了聚类分析树状图。

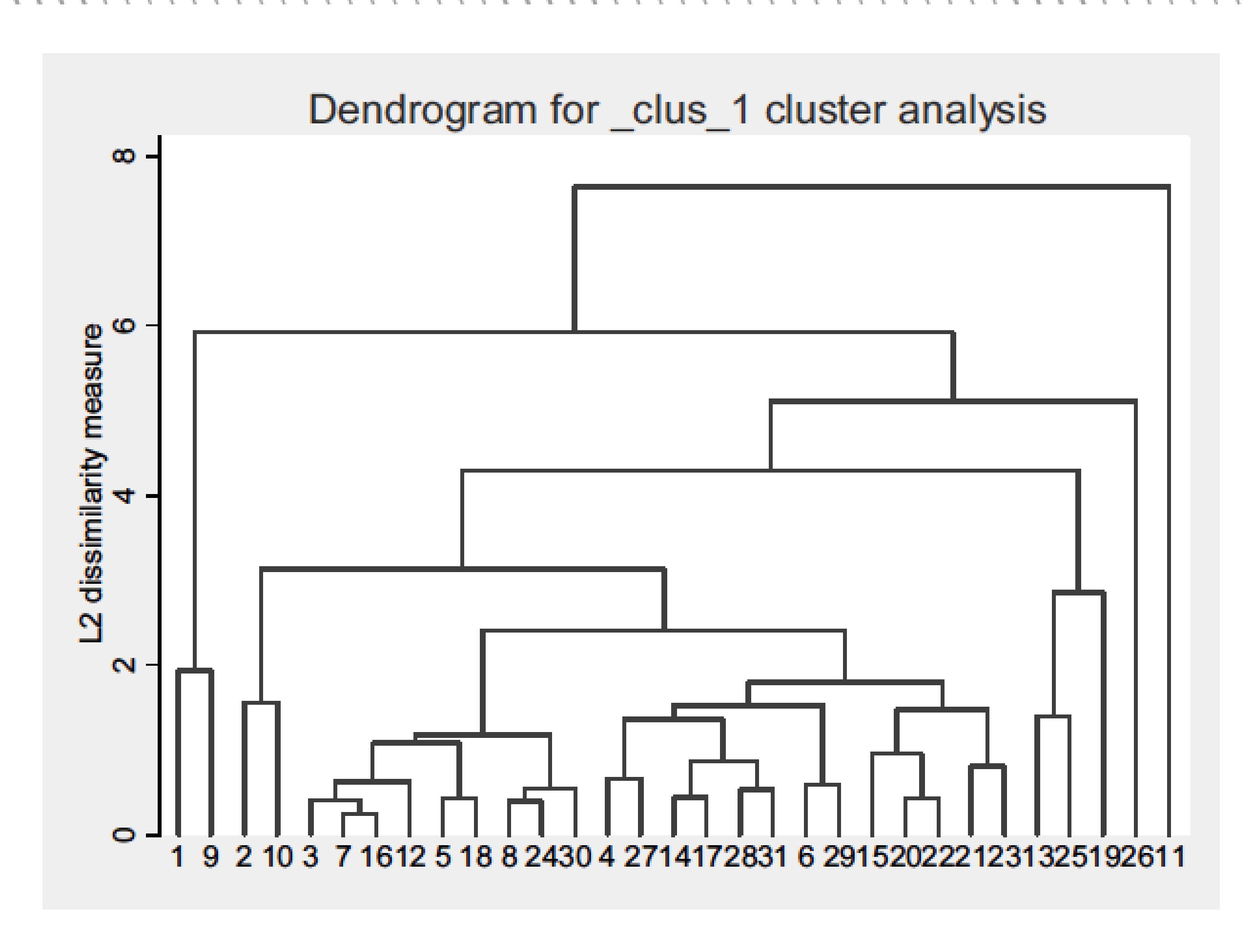


图 9.31　聚类分析树状图

观察图 9.31，可以直观地看到具体的聚类情况：7 号样本与 16 号样本首先聚合在一起，进入数据查看界面查看_clus_1_id 变量，7 号样本代表的是吉林，16 号样本代表的是河南。7 号样本与 16 号样本聚合后又与 3 号样本（河北）聚合，依次类推，最后 11 号样本（浙江）与所有样本聚合为一类。

3．平均联结法聚类分析（Average-Linkage Cluster Analysis）

在分析过程中前 4 条 Stata 命令旨在对数据进行标准化处理，选择的标准化处理方式是使变量的平均数为 0，而且标准差为 1。处理结果与最短联结法聚类分析是一致的，限于篇幅，这里不再赘述。

图 9.32 展示的是使用“平均联结法聚类分析”方法进行分析的结果。在输入第 6 条 Stata 命令并且分别按键盘上的回车键进行确认后，可以看到系统产生了一个新的变量，即聚类变量_clus_1（cluster name: _clus_1）。

```
. cluster averagelinkage zv2 zv3 zv4 zv5
cluster name: _clus_1
```

图 9.32　平均联结法聚类分析结果图

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 9.33 所示的_clus_1 数据。

为了使聚类分析的结果可视化，需要绘制如图 9.34 所示的聚类分析树状图。在输入第 7 条 Stata 命令并且分别按键盘上的回车键进行确认后，可以看到系统产生了聚类分析树状图。

	v1	zv2	zv3	zv4	zv5	_clus_1_id	_clus_1_ord	_clus_1_hgt
1	北京	2.491832	-1.300932	.1381473	2.74791	1	1	1.9333945
2	天津	.2327827	.0430095	-.1851475	1.904427	2	9	4.4272435
3	河北	-.5192236	-.0215061	-.4197636	-.0583391	3	2	1.556306
4	山西	-.2078589	-.8926602	-.2751256	-.5914798	4	10	2.442894
5	内蒙古	-.3863757	.6143994	-.0850018	-.6589216	5	3	.34396065
6	辽宁	-.6192325	-.4909025	-.8162181	.552722	6	7	.24177225
7	吉林	-.7004963	.1837896	-.7007831	-.1394611	7	16	.49672512
8	黑龙江	-.8764872	.6175225	-.7598766	-.2446946	8	12	.65157546
9	上海	2.744534	.4372417	.9158435	2.528599	9	14	.44309643
10	江苏	.3243138	1.123207	.5200769	1.038666	10	17	.56872284
11	浙江	1.782139	3.779561	2.909354	.8466673	11	29	.93349776
12	安徽	-.4867221	-.0903049	-.3652568	-.5288357	12	6	.99642891
13	福建	.4865569	.5600563	2.106966	.7204288	13	28	.52560714
14	江西	-.5702236	-.3525019	-.6970371	-.3460047	14	31	1.1017742
15	山东	.5832218	-.4727882	-.1567856	-.601436	15	4	.65904896
16	河南	-.6224905	.0291783	-.5855773	-.2627164	16	27	1.0187128
17	湖北	-.4244509	-.6880782	-.4589045	-.2700696	17	20	.42418311
18	湖南	-.3305052	.6580344	.177594	-.3206295	18	22	.70076941
19	广东	1.915398	1.167288	1.872044	-.0439652	19	23	1.3086976
20	广西	-.2605374	-.3902475	.0287515	-.208404	20	5	.43407789
21	海南	-.6324823	.0312902	.1970879	-.7896157	21	18	.77219715
22	重庆	.0986308	-.4718364	.121176	-.0193789	22	21	1.0480483
23	四川	-.7282998	-.4032756	.2984568	-.125363	23	8	.39210278
24	贵州	-.8607938	.4157366	-.6277761	-.5534505	24	24	.50638234
25	云南	-.5914577	-.1654393	1.954759	.2183442	25	30	1.4209586
26	西藏	1.148904	-1.808074	-1.23813	-2.248093	26	15	3.0458377
27	陕西	-.5186969	-1.403134	-.2845287	-.313894	27	13	1.4013262
28	甘肃	-.461506	-.824605	-1.015133	-.6959153	28	25	2.2993799
29	青海	-.8206413	-.4094625	-.8431275	.0066801	29	19	3.4891544
30	宁夏	-.7557332	.9089285	-.825086	-.5872713	30	26	5.5369152
31	新疆	-.4340971	-.3834955	-.9009979	-.9565053	31	11	.

图 9.33　_clus_1 数据

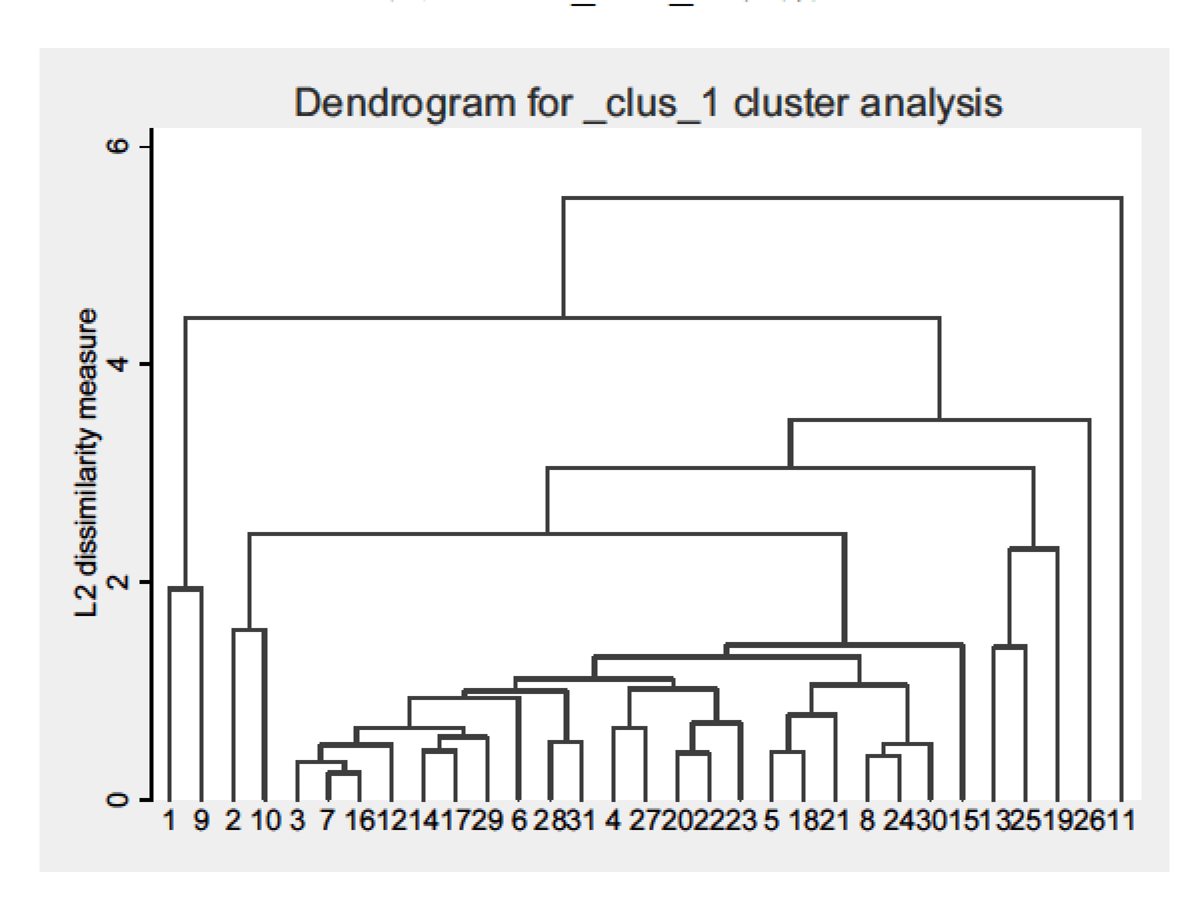


图 9.34　聚类分析树状图

观察图 9.34，可以直观地看到具体的聚类情况：7 号样本与 16 号样本首先聚合在一起，进入数据查看界面查看_clus_1_id 变量， 7 号样本代表的是吉林，16 号样本代表的是河南。7 号样本与 16 号样本聚合后又与 3 号样本（河北）聚合，依次类推，最后 11 号样本（浙江）与所有样本聚合为一类。

4．加权平均联结法聚类分析（Weighted-Average Linkage Cluster Analysis）

在分析过程中前 4 条 Stata 命令旨在对数据进行标准化处理，选择的标准化处理方式是使变量的平均数为 0 而且标准差为 1。处理结果与最短联结法聚类分析是一致的，限于篇幅，这里不再赘述。

图 9.35 展示的是使用“加权平均联结法聚类分析”方法进行分析的结果。在输入第 6 条 Stata 命令并且分别按键盘上的回车键进行确认后，可以看到系统产生了一个新的变量，即聚

类变量_clus_1（cluster name: _clus_1）。

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 9.36 所示的_clus_1 数据。

```
. cluster waveragelinkage  zv2 zv3 zv4 zv5
cluster name:  _clus_1
```

图 9.35　加权平均联结法聚类分析结果图

	V1	zv2	zv3	zv4	zv5	_clus_1_id	_clus_1_ord	_clus_1_hgt
1	北京	2.491832	-1.300932	.1381473	2.74791	1	1	1.9333945
2	天津	.2327827	.0430095	-.1851475	1.904427	2	9	3.7787763
3	河北	-.5192236	-.0215061	-.4197636	-.0583391	3	2	1.556306
4	山西	-.2078589	-.8926602	-.2751256	-.5914798	4	10	2.4456021
5	内蒙古	-.3863757	.6143994	-.0850018	-.6589216	5	3	.34396065
6	辽宁	-.6192325	-.4909025	-.8162181	.552722	6	7	.24177225
7	吉林	-.7004963	.1837896	-.7007831	-.1394611	7	16	.49247288
8	黑龙江	-.8764872	.6175225	-.7598766	-.2446946	8	12	.675385
9	上海	2.744534	.4372417	.9158435	2.528599	9	14	.44309643
10	江苏	.3243138	1.123207	.5200769	1.038666	10	17	.56872284
11	浙江	1.782139	3.779561	2.909354	.8466673	11	29	.91761047
12	安徽	-.4867221	-.0903049	-.3652568	-.5288357	12	28	.52560714
13	福建	.4865569	.5600563	2.106966	.7204288	13	31	1.1271187
14	江西	-.5702236	-.3525019	-.6970371	-.3460047	14	4	.65904896
15	山东	.5832218	-.4727882	-.1567856	-.601436	15	27	1.0414301
16	河南	-.6224905	.0291783	-.5855773	-.2627164	16	20	.42418311
17	湖北	-.4244509	-.6880782	-.4589045	-.2700696	17	22	.70076941
18	湖南	-.3305052	.6580344	.177594	-.3206295	18	23	1.2519968
19	广东	1.915398	1.167288	1.872044	-.0439652	19	6	1.5324006
20	广西	-.2605374	-.3902475	.0287515	-.208404	20	15	1.6267549
21	海南	-.6324823	.0312902	.1970879	-.7896157	21	5	.43407789
22	重庆	.0986308	-.4718364	.121176	-.0193789	22	18	.77219715
23	四川	-.7282998	-.4032756	.2984568	-.125363	23	21	1.1128042
24	贵州	-.8607938	.4157366	-.6277761	-.5534505	24	8	.39210278
25	云南	-.5914577	-.1654393	1.954759	.2183442	25	24	.50638234
26	西藏	1.148904	-1.808074	-1.23813	-2.248093	26	30	2.895477
27	陕西	-.5186969	-1.403134	-.2845287	-.313894	27	13	1.4013262
28	甘肃	-.461506	-.824605	-1.015133	-.6959153	28	25	2.2993799
29	青海	-.8206413	-.4094625	-.8431275	.0066801	29	19	4.7783081
30	宁夏	-.7557332	.9089285	-.825086	-.5872713	30	11	6.3446125
31	新疆	-.4340971	-.3834955	-.9009979	-.9565053	31	26	.

图 9.36　_clus_1 数据

为了使聚类分析的结果可视化，需要绘制如图 9.37 所示的聚类分析树状图。在输入第 7 条 Stata 命令并且分别按键盘上的回车键进行确认后，可以看到系统产生了聚类分析树状图。

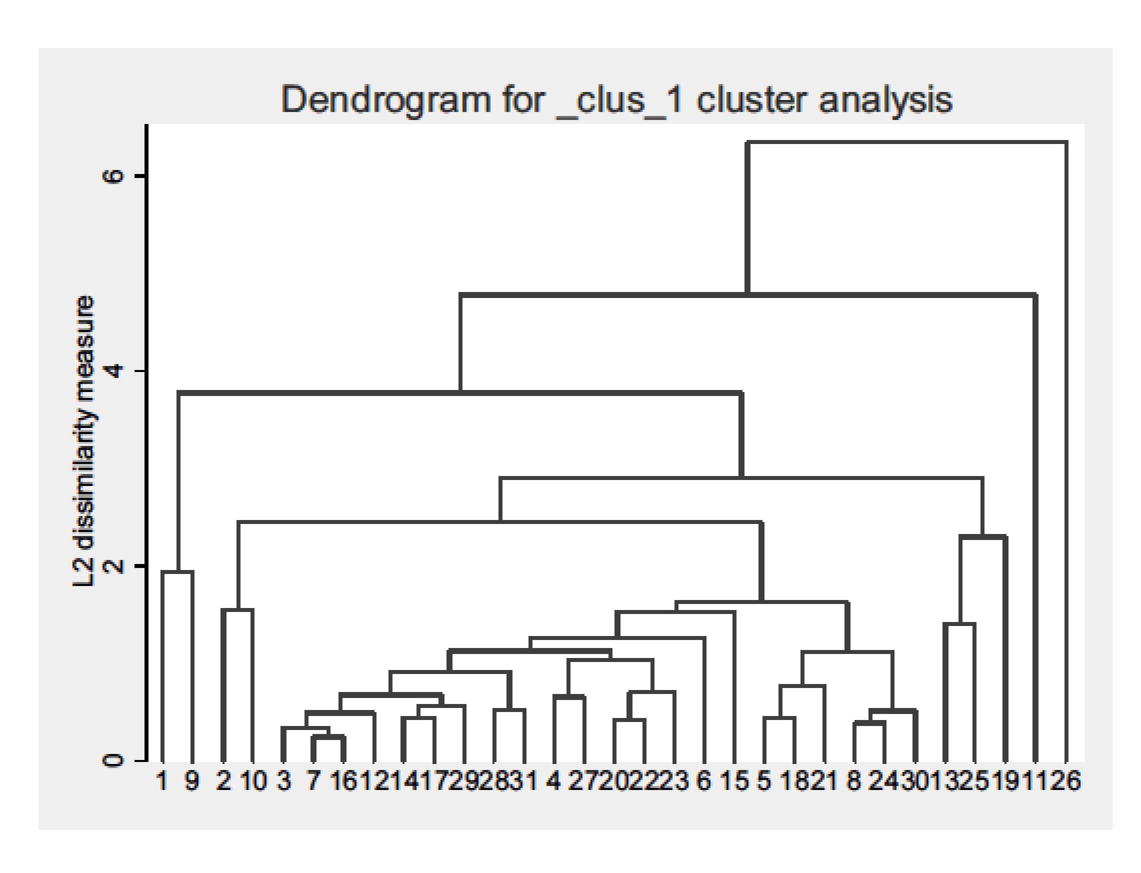


图 9.37　聚类分析树状图

观察图 9.37，可以直观地看到具体的聚类情况：7 号样本与 16 号样本首先聚合在一起，进入数据查看界面查看_clus_1_id 变量， 7 号样本代表的是吉林，16 号样本代表的是河南，7 号样本与 16 号样本聚合后又与 3 号样本（河北）聚合，依次类推。最后，11 号样本（浙江）与所有样本聚合为一类。

5．中位数联结法聚类分析（Median-Linkage Cluster Analysis）

在分析过程中前 4 条 Stata 命令旨在对数据进行标准化处理，选择的标准化处理方式是使

变量的平均数为 0 而且标准差为 1。处理结果与最短联结法聚类分析是一致的，限于篇幅，这里不再赘述。

图 9.38 展示的是使用“中位数联结法聚类分析”方法进行分析的结果。在输入第 6 条 Stata 命令并且分别按键盘上的回车键进行确认后，可以看到系统产生了一个新的变量，即聚类变量_clus_1（cluster name: _clus_1）。

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 9.39 所示的_clus_1 数据。

```
. cluster medianlinkage zv2 zv3 zv4 zv5
cluster name: _clus_1
```

图 9.38　中位数联结法聚类分析结果图

	v1	zv2	zv3	zv4	zv5	_clus_1_id	_clus_1_ord	_clus_1_hgt
1	北京	2.491832	-1.300932	.1381473	2.74791	1	1	3.7380142
2	天津	.2327827	.0430095	-.1851475	1.904427	2	9	11.055998
3	河北	-.5192236	-.0215061	-.4197636	-.0583391	3	2	2.4220884
4	山西	-.2078589	-.8926602	-.2751256	-.5914798	4	10	4.7349776
5	内蒙古	-.3863757	.6143994	-.0850018	-.6589216	5	13	1.963715
6	辽宁	-.6192325	-.4909025	-.8162181	.552722	6	25	4.8921571
7	吉林	-.7004963	.1837896	-.7007831	-.1394611	7	3	.10691459
8	黑龙江	-.8764872	.6175225	-.7598766	-.2446946	8	7	.05845382
9	上海	2.744534	.4372417	.9158435	2.528599	9	16	.21530028
10	江苏	.3243138	1.123207	.5200769	1.038666	10	12	.26534172
11	浙江	1.782139	3.779561	2.909354	.8466673	11	14	.19633444
12	安徽	-.4867221	-.0903049	-.3652568	-.5288357	12	17	.3252146
13	福建	.4865569	.5600563	2.106966	.7204288	13	29	.54637565
14	江西	-.5702236	-.3525019	-.6970371	-.3460047	14	6	.98708902
15	山东	.5832218	-.4727882	-.1567856	-.601436	15	4	.43434553
16	河南	-.6224905	.0291783	-.5855773	-.2627164	16	27	.90244057
17	湖北	-.4244509	-.6880782	-.4589045	-.2700696	17	28	.27626287
18	湖南	-.3305052	.6580344	.177594	-.3206295	18	31	1.0970005
19	广东	1.915398	1.167288	1.872044	-.0439652	19	5	.18842361
20	广西	-.2605374	-.3902475	.0287515	-.208404	20	18	.55367749
21	海南	-.6324823	.0312902	.1970879	-.7896157	21	21	.84681882
22	重庆	.0986308	-.4718364	.121176	-.0193789	22	20	.17993131
23	四川	-.7282998	-.4032756	.2984568	-.125363	23	22	.46990925
24	贵州	-.8607938	.4157366	-.6277761	-.5534505	24	23	1.2654128
25	云南	-.5914577	-.1654393	1.954759	.2183442	25	15	2.2995483
26	西藏	1.148904	-1.808074	-1.27813	-2.248093	26	8	.15374459
27	陕西	-.5186969	-1.403134	-.2845287	-.313894	27	24	.21929439
28	甘肃	-.461506	-.824605	-1.015133	-.6959153	28	30	7.5403606
29	青海	-.8206413	-.4094625	-.8431275	.0066801	29	19	17.909095
30	宁夏	-.7557332	.9089285	-.825086	-.5872713	30	11	35.452412
31	新疆	-.4340971	-.3834955	-.9009979	-.9565053	31	26	.

图 9.39　_clus_1 数据

为了使聚类分析的结果可视化，需要绘制如图 9.40 所示的聚类分析树状图。在输入第 7 条 Stata 命令并且分别按键盘上的回车键进行确认后，可以看到系统产生了聚类分析树状图。

观察图 9.40，可以直观地看到具体的聚类情况：7 号样本与 16 号样本首先聚合在一起。进入数据查看界面查看_clus_1_id 变量，7 号样本代表的是吉林，16 号样本代表的是河南，7 号样本与 16 号样本聚合后又与 3 号样本（河北）聚合，依次类推。最后，11 号样本（浙江）与所有样本聚合为一类。

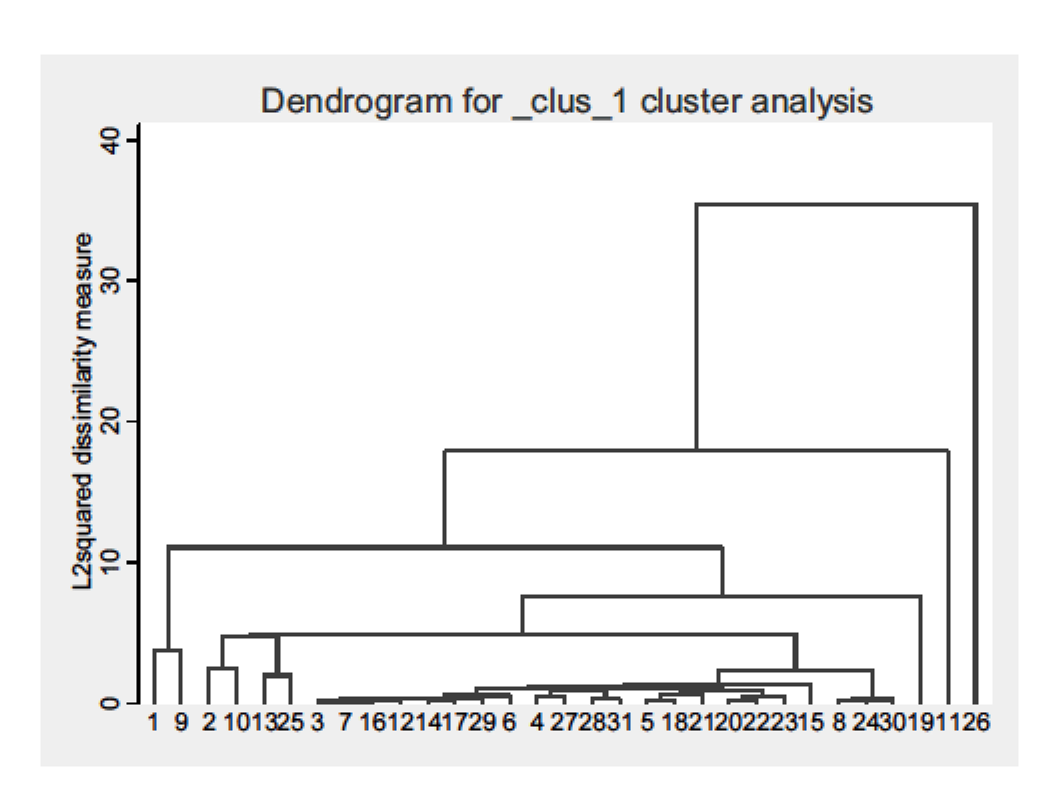


图 9.40　聚类分析树状图

6．重心联结法聚类分析（Centroid-Linkage Cluster Analysis）

在分析过程中前 4 条 Stata 命令旨在对数据进行标准化处理，选择的标准化处理方式是使变量的平均数为 0 而且标准差为 1。处理结果与最短联结法聚类分析是一致的，限于篇幅，这里不再赘述。

图 9.41 展示的是使用“重心联结法聚类分析”方法进行分析的结果。在输入第 6 条 Stata 命令并且分别按键盘上的回车键进行确认后，可以看到系统产生了一个新的变量，即聚类变量_clus_1（cluster name: _clus_1）。

```
. cluster centroidlinkage zv2 zv3 zv4 zv5
cluster name: _clus_1
```

图 9.41　重心联结法聚类分析结果图

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 9.42 所示的_clus_1 数据。

	v1	zv2	zv3	zv4	zv5	_clus_1_id	_clus_1_ord	_clus_1_pht	_clus_1_hgt
1	北京	2.491832	-1.300932	.1381473	2.74791	1	1	24	3.7380142
2	天津	.2327827	.0430095	-.1851475	1.904427	2	9	29	17.509717
3	河北	-.5192236	-.0215061	-.4197636	-.0583391	3	11	27	8.7109641
4	山西	-.2078589	-.8926602	-.2751256	-.5914798	4	19	30	13.799987
5	内蒙古	-.3863757	.6143994	-.0850018	-.6589216	5	2	23	2.4220884
6	辽宁	-.6192325	-.4909025	-.8162181	.552722	6	10	25	4.7349776
7	吉林	-.7004963	.1837896	-.7007831	-.1394611	7	13	22	1.963715
8	黑龙江	-.8764872	.6175225	-.7598766	-.2446946	8	25	26	4.6882858
9	上海	2.744534	.4372417	.9158435	2.528599	9	3	2	.10691459
10	江苏	.3243138	1.123207	.5200769	1.038666	10	7	1	.05845382
11	浙江	1.782139	3.779561	2.909354	.8466673	11	16	8	.22223725
12	安徽	-.4867221	-.0903049	-.3652568	-.5288357	12	12	11	.28324428
13	福建	.4865569	.5600563	2.106966	.7204288	13	14	6	.19633444
14	江西	-.5702236	-.3525019	-.6970371	-.3460047	14	17	10	.28616324
15	山东	.5832218	-.4727882	-.1567856	-.601436	15	29	15	.68622727
16	河南	-.6224905	.0291783	-.5855773	-.2627164	16	20	4	.17993131
17	湖北	-.4244509	-.6880782	-.4589045	-.2700696	17	22	13	.46990925
18	湖南	-.3305052	.6580344	.177594	-.3206295	18	23	16	.83156462
19	广东	1.915398	1.167288	1.872044	-.0439652	19	6	17	.89322408
20	广西	-.2605374	-.3902475	.0287515	-.208404	20	4	12	.43434553
21	海南	-.6324823	.0312902	.1970879	-.7896157	21	27	18	.78756252
22	重庆	.0986308	-.4718364	.121176	-.0193789	22	28	9	.27626287
23	四川	-.7282998	-.4032756	.2984568	-.125363	23	31	20	1.0759919
24	贵州	-.8607938	.4157366	-.6277761	-.5534505	24	5	5	.18842361
25	云南	-.5914577	-.1654393	1.954759	.2183442	25	18	14	.55367749
26	西藏	1.148904	-1.808074	-1.23813	-2.248093	26	21	19	.90277086
27	陕西	-.5186969	-1.403134	-.2845287	-.313894	27	8	3	.15374459
28	甘肃	-.461506	-.824605	-1.015133	-.6959153	28	24	7	.21929439
29	青海	-.8206413	-.4094625	-.8431275	.0066801	29	30	21	1.4745286
30	宁夏	-.7557332	.9089285	-.825086	-.5872713	30	15	28	10.861324
31	新疆	-.4340971	-.3834955	-.9009979	-.9565053	31	26	.	.

图 9.42　_clus_1 数据

与其他的层次聚类分析方法不同的是，重心联结法聚类分析无法绘制树状图。

7．Ward 联结法聚类分析（Ward's Linkage Cluster Analysis）

在分析过程中前 4 条 Stata 命令旨在对数据进行标准化处理，选择的标准化处理方式是使变量的平均数为 0 而且标准差为 1。处理结果与最短联结法聚类分析是一致的，限于篇幅，这里不再赘述。

图 9.43 展示的是使用“Ward 联结法聚类分析”方法进行分析的结果。在输入第 6 条 Stata 命令并且分别按键盘上的回车键进行确认后，可以看到系统产生了一个新的变量，即聚类变量_clus_1（cluster name: _clus_1）。

```
. cluster wardslinkage zv2 zv3 zv4 zv5
cluster name: _clus_1
```

图 9.43　Ward 联结法聚类分析结果图

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 9.44 所示的_clus_1 数据。

v1	zv2	zv3	zv4	zv5	_clus_1_id	_clus_1_ord	_clus_1_hgt	type1	type2
北京	2.491832	-1.300932	.1381473	2.74791	1	1	3.7380142	1	1
天津	.2327827	.0430095	-.1851475	1.904427	2	9	31.400796	2	1
河北	-.5192236	-.0215061	-.4197636	-.0583391	3	2	2.4220884	4	2
山西	-.2078589	-.8926602	-.2751256	-.5914798	4	10	11.433202	4	2
内蒙古	-.3863757	.6143994	-.0850018	-.6589216	5	13	1.963715	4	2
辽宁	-.6192325	-.4909025	-.8162181	.552722	6	25	6.802721	4	2
吉林	-.7004963	.1837896	-.7007831	-.1394611	7	19	24.863574	4	2
黑龙江	-.8764872	.6175225	-.7598766	-.2446946	8	11	109.13053	4	2
上海	2.744534	.4372417	.9158435	2.528599	9	3	.14255279	1	1
江苏	.3243138	1.123207	.5200769	1.038666	10	7	.05845382	2	1
浙江	1.782139	3.779561	2.909354	.8466673	11	16	.33335587	3	1
安徽	-.4867221	-.0903049	-.3652568	-.5288357	12	12	.83243608	4	2
福建	.4865569	.5600563	2.106966	.7204288	13	14	.19633444	2	1
江西	-.5702236	-.3525019	-.6970371	-.3460047	14	17	1.4952316	4	2
山东	.5832218	-.4727882	-.1567856	-.601436	15	6	.34608393	4	2
河南	-.6224905	.0291783	-.5855773	-.2627164	16	29	4.1243867	4	2
湖北	-.4244509	-.6880782	-.4589045	-.2700696	17	4	.43434553	4	2
湖南	-.3305052	.6580344	.177594	-.3206295	18	27	1.8048811	4	2
广东	1.915398	1.167288	1.872044	-.0439652	19	28	.27626287	2	1
广西	-.2605374	-.3902475	.0287515	-.208404	20	31	3.7690631	4	2
海南	-.6324823	.0312902	.1970879	-.7896157	21	15	1.6570483	4	2
重庆	.0986308	-.4718364	.121176	-.0193789	22	20	.17993131	4	2
四川	-.7282998	-.4032756	.2984568	-.125363	23	22	.62654567	4	2
贵州	-.8607938	.4157366	-.6277761	-.5534505	24	23	9.5088675	4	2
云南	-.5914577	-.1654393	1.954759	.2183442	25	5	.18842361	2	1
西藏	1.148904	-1.808074	-1.23813	-2.248093	26	18	.73823665	4	2
陕西	-.5186969	-1.403134	-.2845287	-.313894	27	21	2.7083126	4	2
甘肃	-.461506	-.824605	-1.015133	-.6959153	28	8	.15374459	4	2
青海	-.8206413	-.4094625	-.8431275	.0066801	29	24	.29239253	4	2
宁夏	-.7557332	.9089285	-.825086	-.5872713	30	30	18.378473	4	2
新疆	-.4340971	-.3834955	-.9009979	-.9565053	31	26	.	4	2

图 9.44　_clus_1 数据

在图 9.44 中，可以看到层次聚类分析方法产生的聚类变量的 3 个组成部分：_clus_1_id、_clus_1_ord、_clus_1_hgt。其中_clus_1_id 表示的是系统对该观测样本的初始编号；_clus_1_ord 表示的是系统对该观测样本进行聚类分析处理后的编号；_clus_1_hgt 表示的是系统对该观测样本进行聚类计算后的值。

为了使聚类分析的结果可视化，需要绘制如图 9.45 所示的聚类分析树状图。在输入第 7 条 Stata 命令并且分别按键盘上的回车键进行确认后，可以看到系统产生了聚类分析树状图。

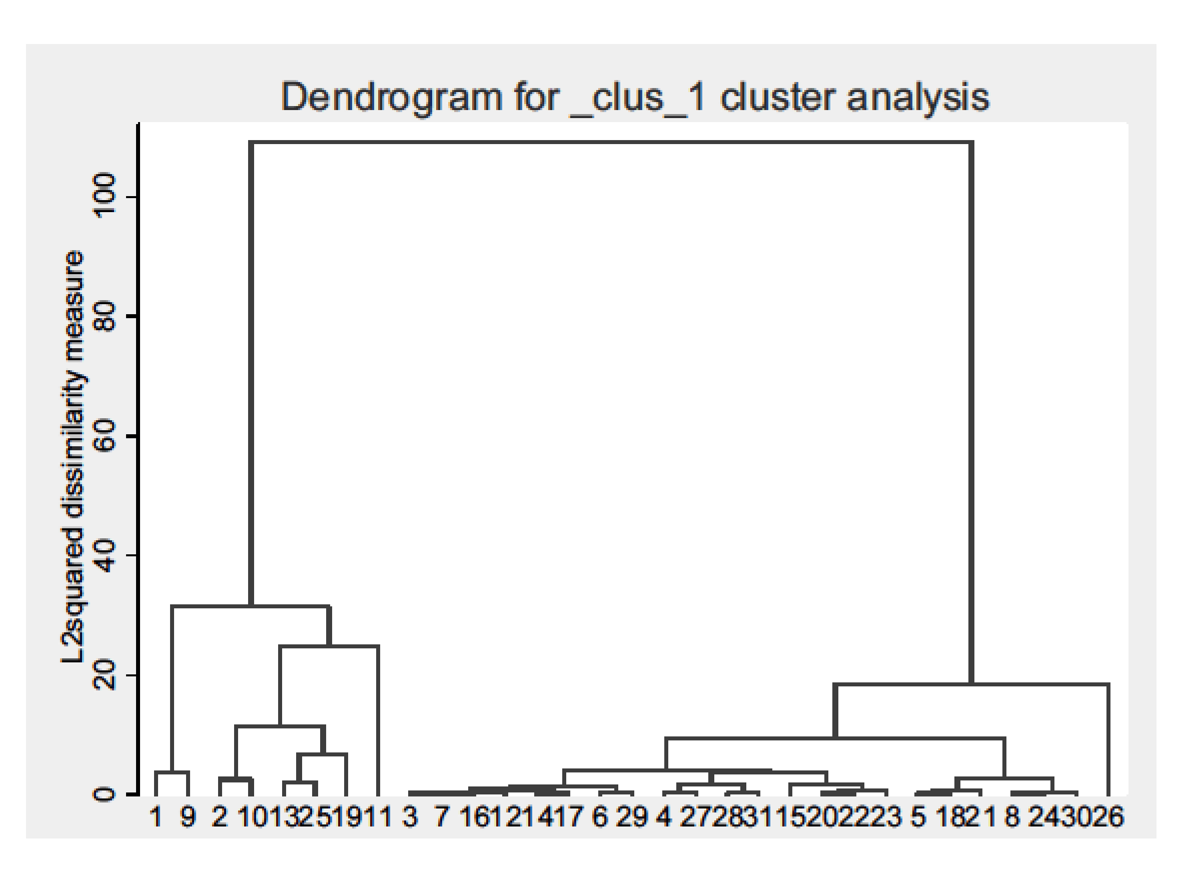


图 9.45　聚类分析树状图

观察图 9.45，可以直观地看到具体的聚类情况：7 号样本与 3 号样本首先聚合在一起，进入数据查看界面查看_clus_1_id 变量，7 号样本代表的是吉林，3 号样本代表的是河北，7 号样本与 3 号样本聚合后又与 16 号样本（河南）聚合，依次类推。

9.2.5 案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

下面将根据拟分类数进行聚类的案例延伸分析。

在以上各种层次聚类分析方法中，如果样本比较多，可能图中就显得比较乱，可以使用产生聚类变量的方法对样本进行有拟分类数的聚类。例如，分别把所有观测样本分为 4 类和 2 类。

操作命令如下。

- cluster generate type1=group(4)：本命令的含义是产生聚类变量 type1，使用层次聚类分析方法，把样本分为 4 类。
- cluster generate type2=group(2)：本命令的含义是产生聚类变量 type2，使用层次聚类分析方法，把样本分为 2 类。

本操作命令对所有层次聚类分析方法均适用。

使用各种层次聚类分析方法对观测样本进行拟分类数的聚类结果如图 9.46~图 9.59 所示。

1．最短联结法聚类分析

图 9.46 展示的是设定聚类数为 4，然后进行分析的结果。在输入第 1 条 Stata 命令并且分别按键盘上的回车键进行确认后，选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 9.46 所示的 type1 数据。

在图 9.46 中，可以看到所有的观测样本被分为 4 类：其中，浙江被分到第 1 类，上海、北京为第 2 类，西藏为第 3 类，其他省市为第 4 类。可以发现第 1 类的特征是经营净收入、财产性收入高；第 2 类的特征是工薪收入、转移性收入高；第 3 类的特征是收入水平普遍较低；第 4 类的特征是所有收入都处在中间水平。

	v1	zv2	zv3	zv4	zv5	_clus_1_id	_clus_1_ord	_clus_1_hgt	type1	type2
1	北京	2.491832	-1.300932	.1381473	2.74791	1	11	2.9514342	2	2
2	天津	.2327827	.0430095	-.1851475	1.904427	2	1	1.9333945	4	2
3	河北	-.5192236	-.0215061	-.4197636	-.0583391	3	9	2.7794725	4	2
4	山西	-.2078589	-.8926602	-.2751256	-.5914798	4	26	2.4461742	4	2
5	内蒙古	-.3863757	.6143994	-.0850018	-.6589216	5	19	1.7463687	4	2
6	辽宁	-.6192325	-.4909025	-.8162181	.552722	6	13	1.4013262	4	2
7	吉林	-.7004963	.1837896	-.7007831	-.1394611	7	25	1.7136991	4	2
8	黑龙江	-.8764872	.6175225	-.7598766	-.2446946	8	2	1.556306	4	2
9	上海	2.744534	.4372417	.9158435	2.528599	9	10	1.6155965	2	2
10	江苏	.3243138	1.123207	.5200769	1.038666	10	15	.80677284	4	2
11	浙江	1.782139	3.779561	2.909354	.8466673	11	5	.43407789	1	1
12	安徽	-.4867221	-.0903049	-.3652568	-.5288357	12	18	.70515268	4	2
13	福建	.4865569	.5600563	2.106966	.7204288	13	27	.65904896	4	2
14	江西	-.5702236	-.3525019	-.6970371	-.3460047	14	21	.64828178	4	2
15	山东	.5832218	-.4727882	-.1567856	-.601436	15	20	.42418311	4	2
16	河南	-.6224906	.0291783	-.5855773	-.2627164	16	22	.54645054	4	2
17	湖北	-.4244509	-.6880782	-.4589045	-.2700696	17	23	.59764691	4	2
18	湖南	-.3305052	.6580344	.177594	-.3206295	18	6	.58828898	4	2
19	广东	1.915398	1.167288	1.872044	-.0439652	19	4	.47523163	4	2
20	广西	-.2605374	-.3902475	.0287515	-.208404	20	14	.40959957	4	2
21	海南	-.6324823	.0312902	.1970879	-.7896157	21	12	.38996133	4	2
22	重庆	.0986308	-.4718364	.121176	-.0193789	22	3	.28722333	4	2
23	四川	-.7282998	-.4032756	.2984568	-.125363	23	7	.24177225	4	2
24	贵州	-.8607938	.4157366	-.6277761	-.5534505	24	16	.44309643	4	2
25	云南	-.5914577	-.1654393	1.954759	.2183442	25	17	.46008964	4	2
26	西藏	1.148904	-1.808074	-1.23813	-2.248093	26	29	.48338714	3	2
27	陕西	-.5186969	-1.403134	-.2845287	-.313894	27	8	.39210278	4	2
28	甘肃	-.461506	-.824605	-1.015133	-.6959153	28	24	.47022339	4	2
29	青海	-.8206413	-.4094625	-.8431275	.0066801	29	30	.65863648	4	2
30	宁夏	-.7557332	.9089285	-.825086	-.5872713	30	28	.52560714	4	2
31	新疆	-.4340971	-.3834955	-.9009979	-.9565053	31	31	.	4	2

图 9.46　最短联结法聚类分析 type1 数据

图 9.47 展示的是设定聚类数为 2，然后进行分析的结果。在输入第 2 条 Stata 命令并且分别按键盘上的回车键进行确认后，选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，

进入数据查看界面，可以看到如图 9.47 所示的 type2 数据。

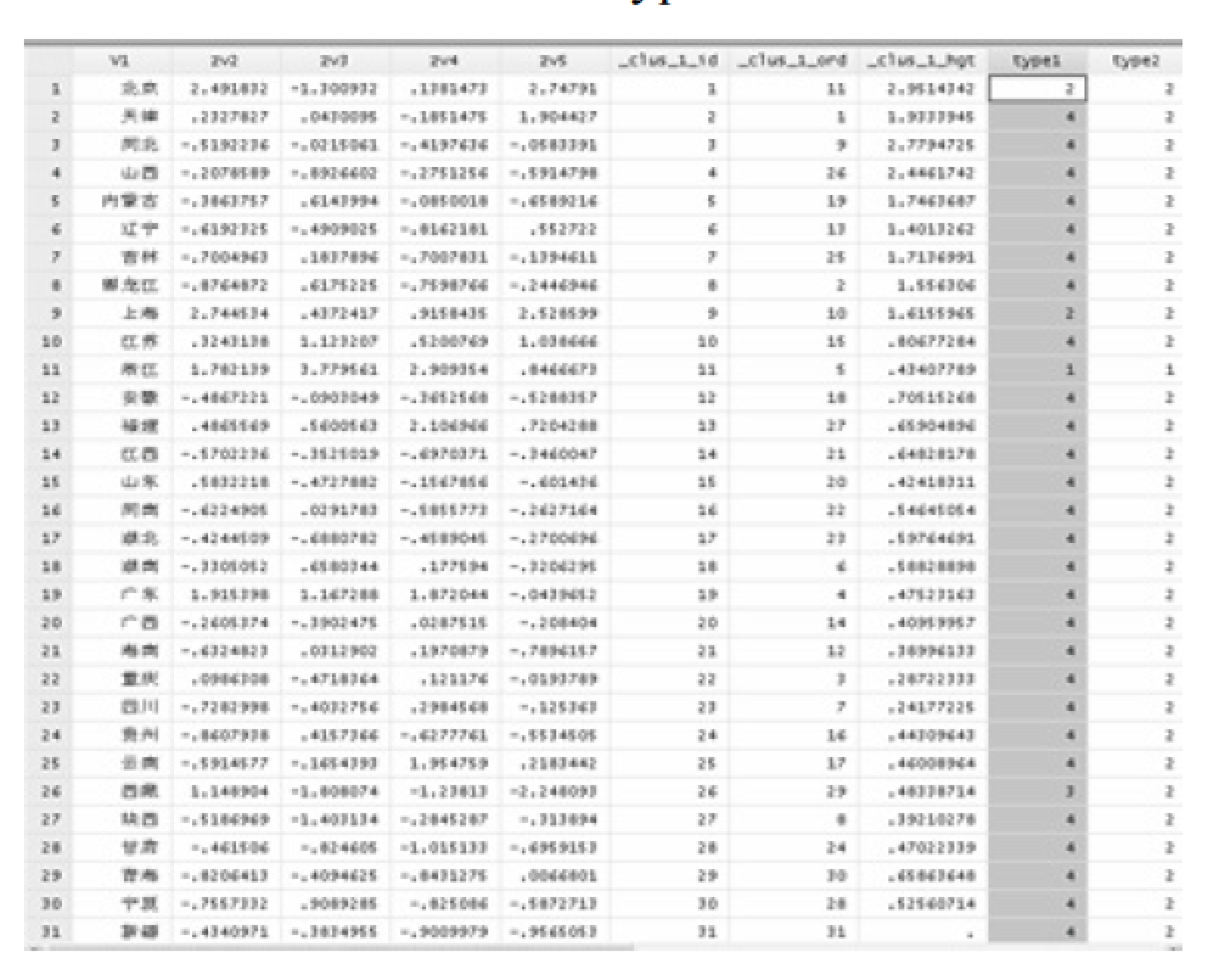
	v1	zv2	zv3	zv4	zv5	_clus_1_id	_clus_1_ord	_clus_1_hgt	type1	type2
1	北京	2.491832	-1.300932	.1381473	2.74791	1	11	2.9514342	2	2
2	天津	.2327827	.0430095	-.1851475	1.904427	2	1	1.9333945	4	2
3	河北	-.5192236	-.0215061	-.4197636	-.0583391	3	9	2.7794725	4	2
4	山西	-.2078589	-.8926602	-.2751256	-.5914798	4	26	2.4461742	4	2
5	内蒙古	-.3863757	.6143994	-.0850018	-.6589216	5	19	1.7463687	4	2
6	辽宁	-.6192325	-.4909025	-.8162181	.552722	6	13	1.4013262	4	2
7	吉林	-.7004963	.1837896	-.7007831	-.1394611	7	25	1.7136991	4	2
8	黑龙江	-.8764872	.6175225	-.7598766	-.2446946	8	2	1.556306	4	2
9	上海	2.744534	.4372417	.9158435	2.528599	9	10	1.6155965	2	2
10	江苏	.3243138	1.123207	.5200769	1.038666	10	15	.80677284	4	2
11	浙江	1.782139	3.779561	2.909354	.8466673	11	5	.43407789	1	1
12	安徽	-.4867221	-.0903049	-.3652568	-.5288357	12	18	.70515268	4	2
13	福建	.4865569	.5600563	2.106966	.7204288	13	27	.65904896	4	2
14	江西	-.5702236	-.3525019	-.6970371	-.3460047	14	21	.64828178	4	2
15	山东	.5832218	-.4727882	-.1567856	-.601436	15	20	.42418311	4	2
16	河南	-.6224905	.0291783	-.5855773	-.2627164	16	22	.54645054	4	2
17	湖北	-.4244509	-.6880782	-.4589045	-.2700696	17	23	.59764691	4	2
18	湖南	-.3305052	.6580344	.177594	-.3206295	18	6	.58828898	4	2
19	广东	1.915398	1.167288	1.872044	-.0439652	19	4	.47523163	4	2
20	广西	-.2605374	-.3902475	.0287515	-.208404	20	14	.40959957	4	2
21	海南	-.6324823	.0312902	.1970879	-.7896157	21	12	.38996133	4	2
22	重庆	.0986308	-.4718364	.121176	-.0193789	22	3	.28722333	4	2
23	四川	-.7282998	-.4032756	.2984568	-.125363	23	7	.24177225	4	2
24	贵州	-.8607938	.4157366	-.6277761	-.5534505	24	16	.44309643	4	2
25	云南	-.5914577	-.1654393	1.954759	.2183442	25	17	.46008964	4	2
26	西藏	1.148904	-1.808074	-1.23813	-2.248093	26	29	.48338714	3	2
27	陕西	-.5186969	-1.403134	-.2845287	-.313894	27	8	.39210278	4	2
28	甘肃	-.461506	-.824605	-1.015133	-.6959153	28	24	.47022339	4	2
29	青海	-.8206413	-.4094625	-.8431275	.0066801	29	30	.65863648	4	2
30	宁夏	-.7557332	.9089285	-.825086	-.5872713	30	28	.52560714	4	2
31	新疆	-.4340971	-.3834955	-.9009979	-.9565053	31	31	.	4	2

图 9.47　最短联结法聚类分析 type2 数据

在图 9.47 中，可以看到所有的观测样本被分为两类。其中，浙江被分到第 1 类，其他省市为第 2 类。第 1 类的特征是经营净收入、财产性收入高；第 2 类的特征不明显。

2．最长联结法聚类分析

图 9.48 展示的是设定聚类数为 4，然后进行分析的结果。在输入第 1 条 Stata 命令并且分别按键盘上的回车键进行确认后，选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 9.48 所示的 type1 数据。

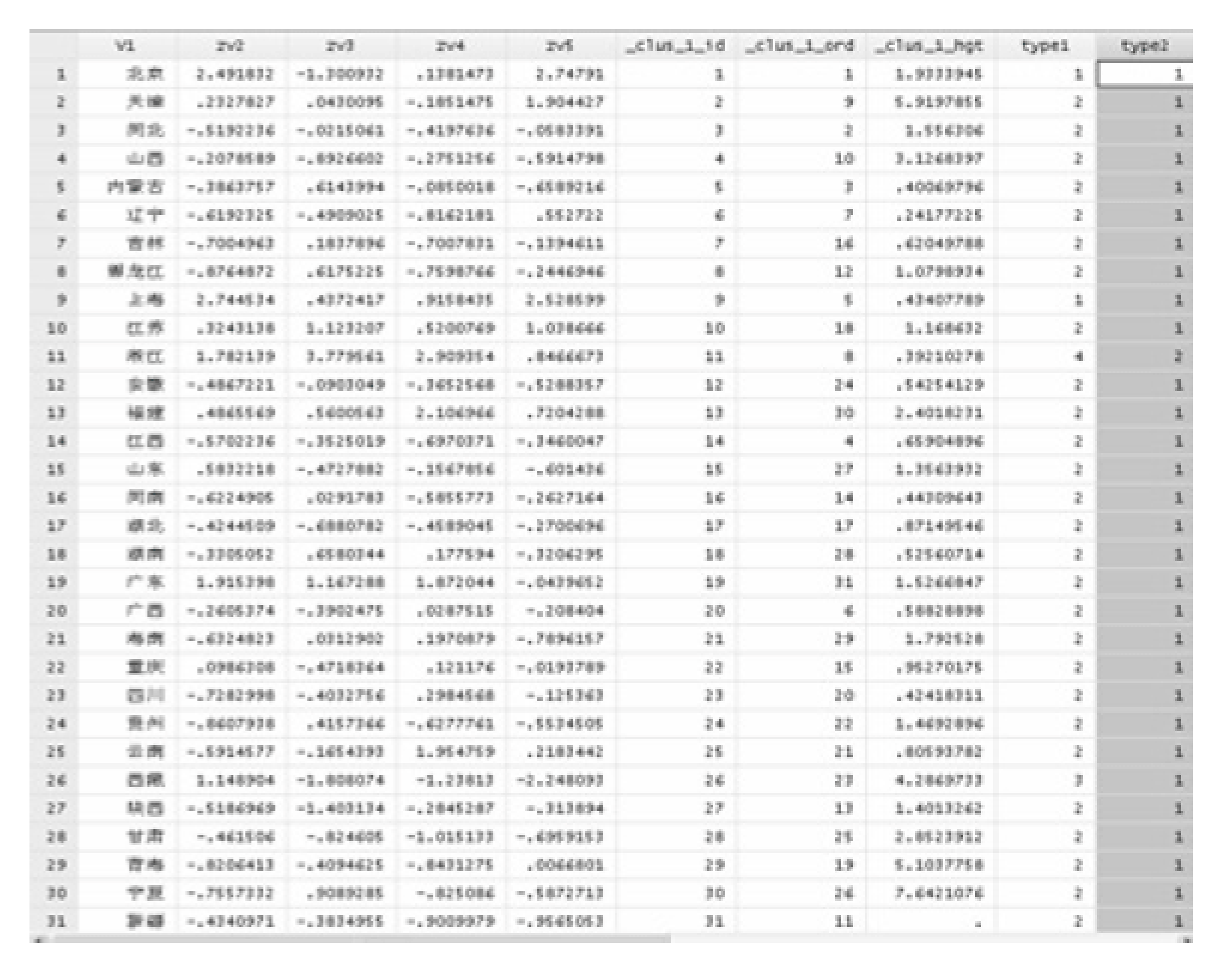


	v1	zv2	zv3	zv4	zv5	_clus_1_id	_clus_1_ord	_clus_1_hgt	type1	type2
1	北京	2.491832	-1.300932	.1381473	2.74791	1	11	2.9514342	2	2
2	天津	.2327827	.0430095	-.1851475	1.904427	2	1	1.9333945	4	2
3	河北	-.5192236	-.0215061	-.4197636	-.0583391	3	9	2.7794725	4	2
4	山西	-.2078589	-.8926602	-.2751256	-.5914798	4	26	2.4461742	4	2
5	内蒙古	-.3863757	.6143994	-.0850018	-.6589216	5	19	1.7463687	4	2
6	辽宁	-.6192325	-.4909025	-.8162181	.552722	6	13	1.4013262	4	2
7	吉林	-.7004963	.1837896	-.7007831	-.1394611	7	25	1.7136991	4	2
8	黑龙江	-.8764872	.6175225	-.7598766	-.2446946	8	2	1.556306	4	2
9	上海	2.744534	.4372417	.9158435	2.528599	9	10	1.6155965	2	2
10	江苏	.3243138	1.123207	.5200769	1.038666	10	15	.80677284	4	2
11	浙江	1.782139	3.779561	2.909354	.8466673	11	5	.43407789	1	1
12	安徽	-.4867221	-.0903049	-.3652568	-.5288357	12	18	.70515268	4	2
13	福建	.4865569	.5600563	2.106966	.7204288	13	27	.65904896	4	2
14	江西	-.5702236	-.3525019	-.6970371	-.3460047	14	21	.64828178	4	2
15	山东	.5832218	-.4727882	-.1567856	-.601436	15	20	.42418311	4	2
16	河南	-.6224905	.0291783	-.5855773	-.2627164	16	22	.54645054	4	2
17	湖北	-.4244509	-.6880782	-.4589045	-.2700696	17	23	.59764691	4	2
18	湖南	-.3305052	.6580344	.177594	-.3206295	18	6	.58828898	4	2
19	广东	1.915398	1.167288	1.872044	-.0439652	19	4	.47523163	4	2
20	广西	-.2605374	-.3902475	.0287515	-.208404	20	14	.40959957	4	2
21	海南	-.6324823	.0312902	.1970879	-.7896157	21	12	.38996133	4	2
22	重庆	.0986308	-.4718364	.121176	-.0193789	22	3	.28722333	4	2
23	四川	-.7282998	-.4032756	.2984568	-.125363	23	7	.24177225	4	2
24	贵州	-.8607938	.4157366	-.6277761	-.5534505	24	16	.44309643	4	2
25	云南	-.5914577	-.1654393	1.954759	.2183442	25	17	.46008964	4	2
26	西藏	1.148904	-1.808074	-1.23813	-2.248093	26	29	.48338714	3	2
27	陕西	-.5186969	-1.403134	-.2845287	-.313894	27	8	.39210278	4	2
28	甘肃	-.461506	-.824605	-1.015133	-.6959153	28	24	.47022339	4	2
29	青海	-.8206413	-.4094625	-.8431275	.0066801	29	30	.65863648	4	2
30	宁夏	-.7557332	.9089285	-.825086	-.5872713	30	28	.52560714	4	2
31	新疆	-.4340971	-.3834955	-.9009979	-.9565053	31	31	.	4	2

图 9.48　最长联结法聚类分析 type1 数据

在图 9.48 中，可以看到所有的观测样本被分为 4 类：其中，浙江被分到第 1 类，上海、北京为第 2 类，西藏为第 3 类，其他省市为第 4 类。第 1 类的特征是经营净收入、财产性收入高；第 2 类的特征是工薪收入、转移性收入高；第 3 类的特征是收入水平普遍较低；第 4 类的特征是所有收入都处在中间水平。处理结果与最短联结法聚类分析是一致的。

图 9.49 展示的是设定聚类数为 2，然后进行分析的结果。在输入第 2 条 Stata 命令并且分别按键盘上的回车键进行确认后，选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 9.49 所示的 type2 数据。



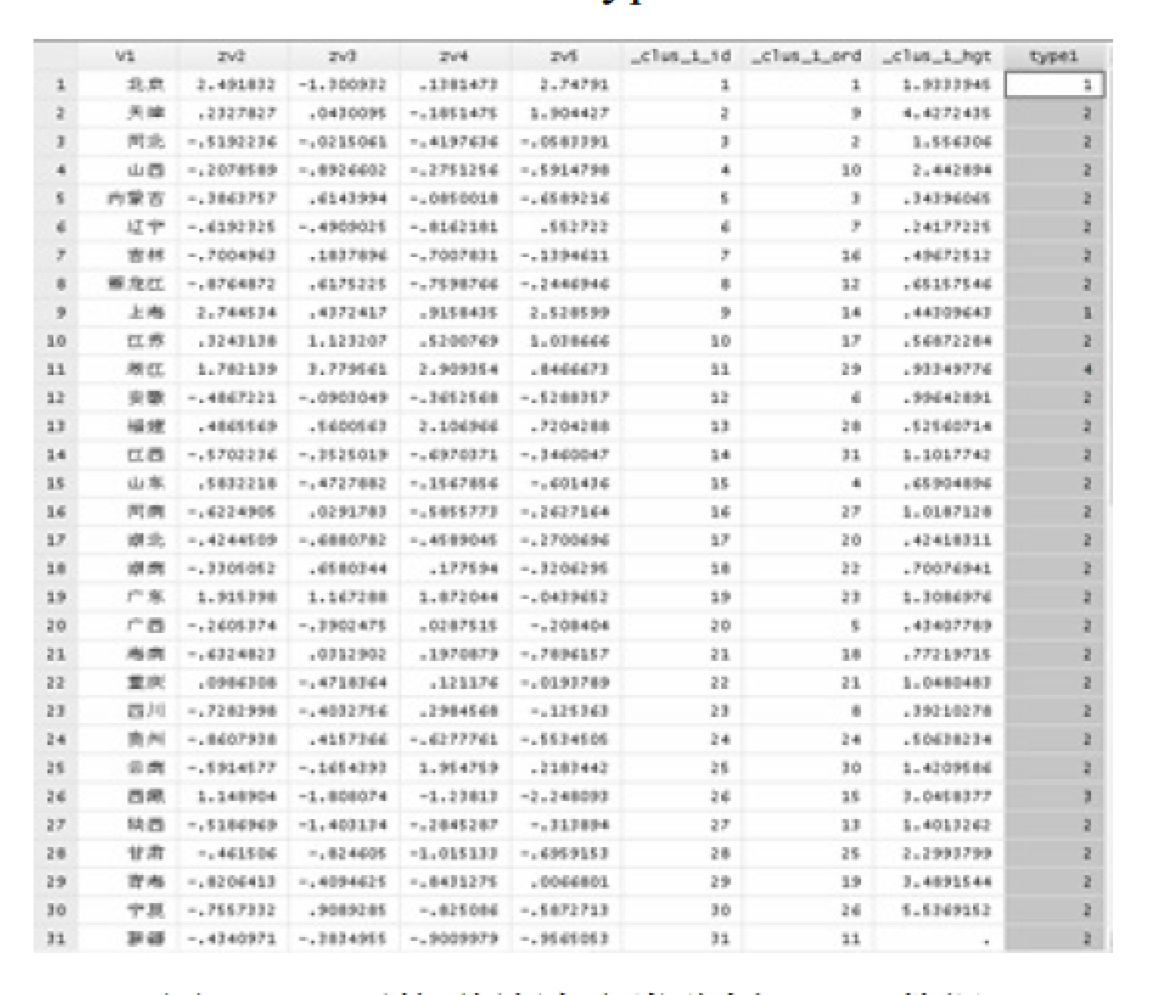
	v1	zv2	zv3	zv4	zv5	_clus_1_id	_clus_1_ord	_clus_1_hgt	type1	type2
1	北京	2.491832	-1.300932	.1381473	2.74791	1	1	1.9333945	1	1
2	天津	.2327827	.0430095	-.1851475	1.904427	2	9	5.9197855	2	1
3	河北	-.5192236	-.0215061	-.4197636	-.0583391	3	2	1.556306	2	1
4	山西	-.2078589	-.8926602	-.2751256	-.5914798	4	10	3.1268397	2	1
5	内蒙古	-.3863757	.6143994	-.0850018	-.6589216	5	3	.40069796	2	1
6	辽宁	-.6192325	-.4909025	-.8162181	.552722	6	7	.24177225	2	1
7	吉林	-.7004963	.1837896	-.7007831	-.1394611	7	16	.62049788	2	1
8	黑龙江	-.8764872	.6175225	-.7598766	-.2446946	8	12	1.0798934	2	1
9	上海	2.744534	.4372417	.9158435	2.528599	9	5	.43407789	1	1
10	江苏	.3243138	1.123207	.5200769	1.038666	10	18	1.168632	2	1
11	浙江	1.782139	3.779561	2.909354	.8466673	11	8	.39210278	4	2
12	安徽	-.4867221	-.0903049	-.3652568	-.5288357	12	24	.54254129	2	1
13	福建	.4865569	.5600563	2.106966	.7204288	13	30	2.4018231	2	1
14	江西	-.5702236	-.3525019	-.6970371	-.3460047	14	4	.65904896	2	1
15	山东	.5832218	-.4727882	-.1567856	-.601436	15	27	1.3563932	2	1
16	河南	-.6224905	.0291783	-.5855773	-.2627164	16	14	.44309643	2	1
17	湖北	-.4244509	-.6880782	-.4589045	-.2700696	17	17	.87149546	2	1
18	湖南	-.3305052	.6580344	.177594	-.3206295	18	28	.52560714	2	1
19	广东	1.915398	1.167288	1.872044	-.0439652	19	31	1.5266847	2	1
20	广西	-.2605374	-.3902475	.0287515	-.208404	20	6	.58828898	2	1
21	海南	-.6324823	.0312902	.1970879	-.7896157	21	29	1.792528	2	1
22	重庆	.0986308	-.4718364	.121176	-.0193789	22	15	.95270175	2	1
23	四川	-.7282998	-.4032756	.2984568	-.125363	23	20	.42418311	2	1
24	贵州	-.8607938	.4157366	-.6277761	-.5534505	24	22	1.4692896	2	1
25	云南	-.5914577	-.1654393	1.954759	.2183442	25	21	.80593782	2	1
26	西藏	1.148904	-1.808074	-1.23813	-2.248093	26	23	4.2869733	3	1
27	陕西	-.5186969	-1.403134	-.2845287	-.313894	27	13	1.4013262	2	1
28	甘肃	-.461506	-.824605	-1.015133	-.6959153	28	25	2.0523912	2	1
29	青海	-.8206413	-.4094625	-.8431275	.0066801	29	19	5.1037758	2	1
30	宁夏	-.7557332	.9089285	-.825086	-.5872713	30	26	7.6421076	2	1
31	新疆	-.4340971	-.3834955	-.9009979	-.9565053	31	11	.	2	1

图 9.49　最长联结法聚类分析 type2 数据

在图 9.49 中，可以看到所有的观测样本被分为两类。其中，浙江被分到第 2 类，其他省市为第 1 类。第 2 类的特征是经营净收入、财产性收入高，第 1 类的特征不明显。处理结果与最短联结法聚类分析是一致的。

3．平均联结法聚类分析

图 9.50 展示的是设定聚类数为 4，然后进行分析的结果。在输入第 1 条 Stata 命令并且分别按键盘上的回车键进行确认后，选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 9.50 所示的 type1 数据。



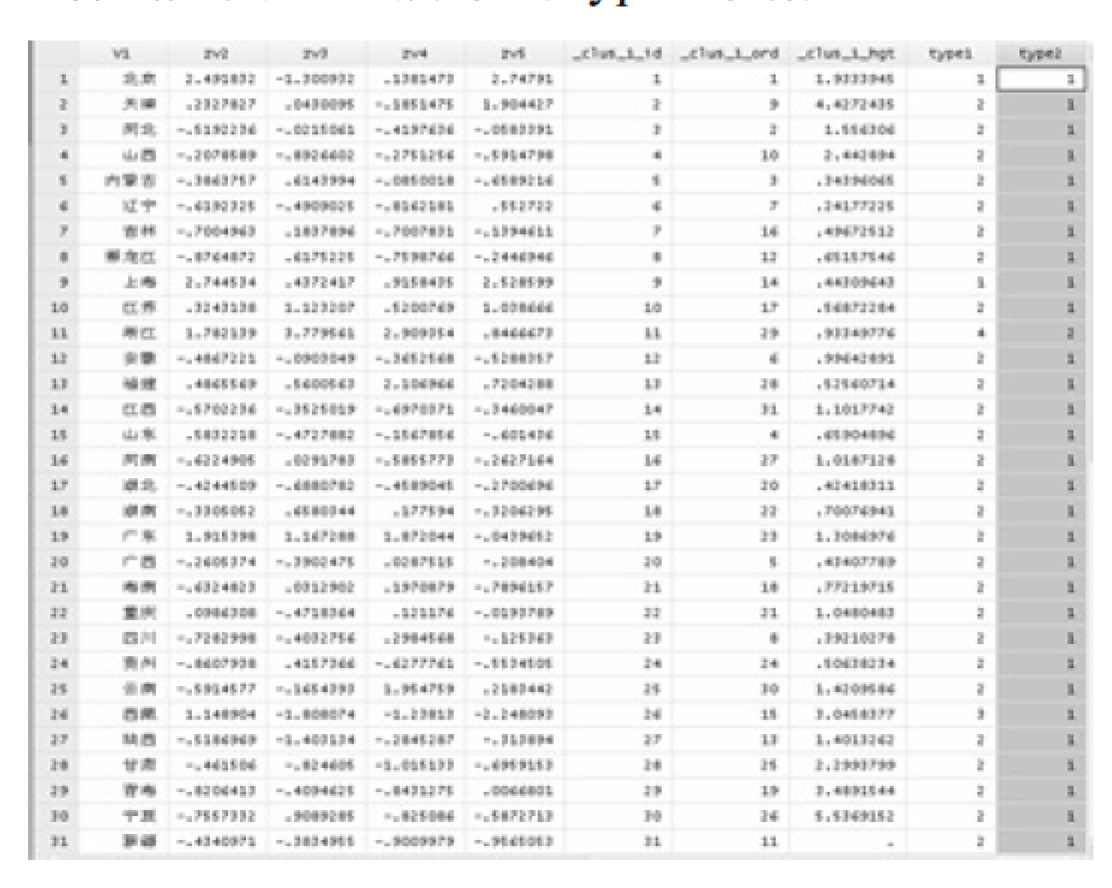
	v1	zv2	zv3	zv4	zv5	_clus_1_id	_clus_1_ord	_clus_1_hgt	type1
1	北京	2.491832	-1.300932	.1381473	2.74791	1	1	1.9333945	1
2	天津	.2327827	.0430095	-.1851475	1.904427	2	9	4.4272435	2
3	河北	-.5192236	-.0215061	-.4197636	-.0583391	3	2	1.556306	2
4	山西	-.2078589	-.8926602	-.2751256	-.5914798	4	10	2.442894	2
5	内蒙古	-.3863757	.6143994	-.0850018	-.6589216	5	3	.34396065	2
6	辽宁	-.6192325	-.4909025	-.8162181	.552722	6	7	.24177225	2
7	吉林	-.7004963	.1837896	-.7007831	-.1394611	7	16	.49672512	2
8	黑龙江	-.8764872	.6175225	-.7598766	-.2446946	8	12	.65157546	2
9	上海	2.744534	.4372417	.9158435	2.528599	9	14	.44309643	1
10	江苏	.3243138	1.123207	.5200769	1.038666	10	17	.56872284	2
11	浙江	1.782139	3.779561	2.909354	.8466673	11	29	.93349776	4
12	安徽	-.4867221	-.0903049	-.3652568	-.5288357	12	6	.99642891	2
13	福建	.4865569	.5600563	2.106966	.7204288	13	28	.52560714	2
14	江西	-.5702236	-.3525019	-.6970371	-.3460047	14	31	1.1017742	2
15	山东	.5832218	-.4727882	-.1567856	-.601436	15	4	.65904896	2
16	河南	-.6224905	.0291783	-.5855773	-.2627164	16	27	1.0187128	2
17	湖北	-.4244509	-.6880782	-.4589045	-.2700696	17	20	.42418311	2
18	湖南	-.3305052	.6580344	.177594	-.3206295	18	22	.70076941	2
19	广东	1.915398	1.167288	1.872044	-.0439652	19	23	1.3086976	2
20	广西	-.2605374	-.3902475	.0287515	-.208404	20	5	.43407789	2
21	海南	-.6324823	.0312902	.1970879	-.7896157	21	18	.77219715	2
22	重庆	.0986308	-.4718364	.121176	-.0193789	22	21	1.0480483	2
23	四川	-.7282998	-.4032756	.2984568	-.125363	23	8	.39210278	2
24	贵州	-.8607938	.4157366	-.6277761	-.5534505	24	24	.50638234	2
25	云南	-.5914577	-.1654393	1.954759	.2183442	25	30	1.4209586	2
26	西藏	1.148904	-1.808074	-1.23813	-2.248093	26	15	3.0458377	3
27	陕西	-.5186969	-1.403134	-.2845287	-.313894	27	13	1.4013262	2
28	甘肃	-.461506	-.824605	-1.015133	-.6959153	28	25	2.2993799	2
29	青海	-.8206413	-.4094625	-.8431275	.0066801	29	19	3.4891544	2
30	宁夏	-.7557332	.9089285	-.825086	-.5872713	30	26	5.5369152	2
31	新疆	-.4340971	-.3834955	-.9009979	-.9565053	31	11	.	2

图 9.50　平均联结法聚类分析 type1 数据

在图 9.50 中，可以看到所有的观测样本被分为 4 类：其中，浙江被分到第 4 类，上海、北京为第 1 类，西藏为第 3 类，其他省市为第 4 类。第 4 类的特征是经营净收入、财产性收入

高。第 1 类的特征是工薪收入、转移性收入高，第 3 类的特征是收入水平普遍较低，第 2 类的特征是所有收入都处在中间水平。处理结果与最短联结法聚类分析是一致的。

图 9.51 展示的是设定聚类数为 2，然后进行分析的结果。在输入第 2 条 Stata 命令并且分别按键盘上的回车键进行确认后，选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 9.51 所示的 type2 数据。



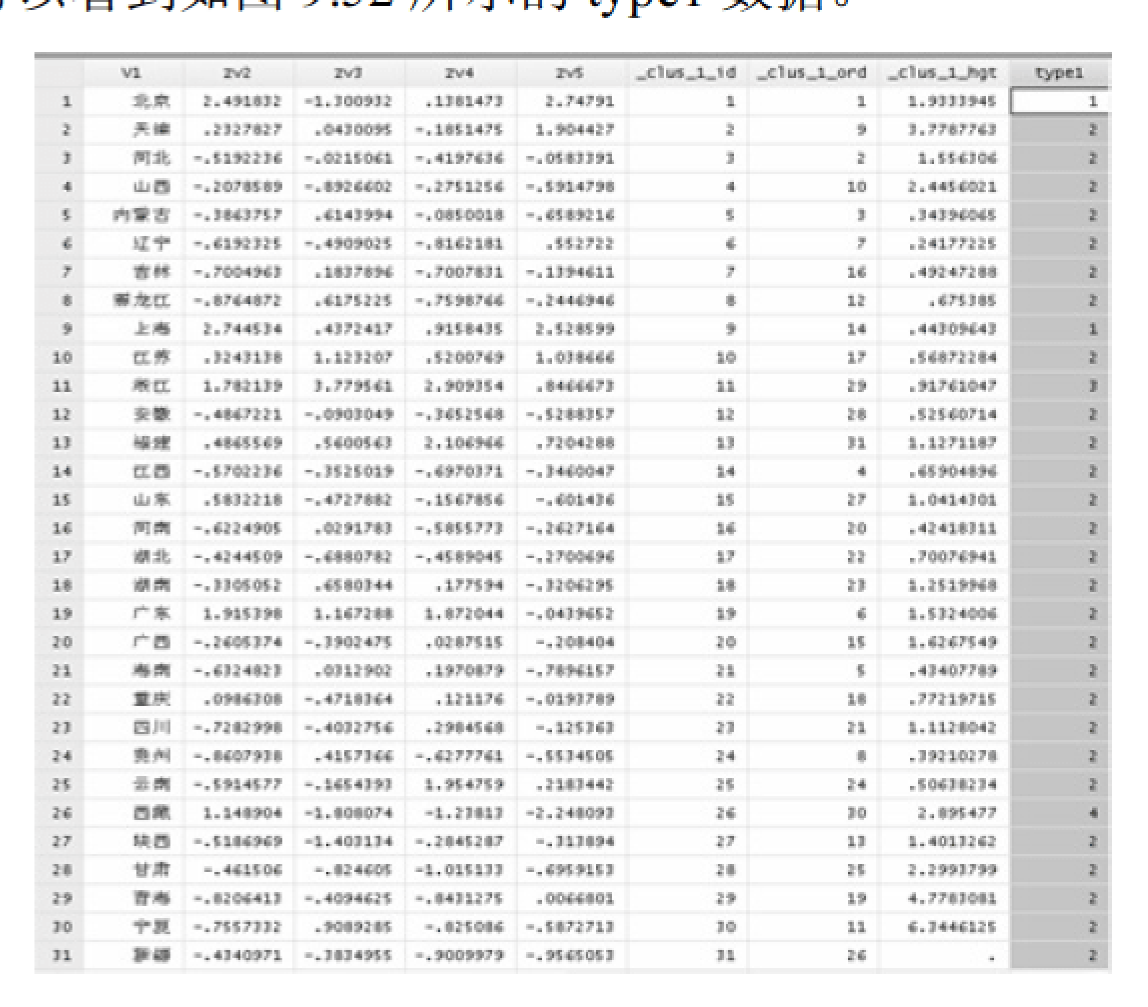
	v1	zv2	zv3	zv4	zv5	_clus_1_id	_clus_1_ord	_clus_1_hgt	type1	type2
1	北京	2.491832	-1.300932	.1381473	2.74791	1	1	1.9333945	1	1
2	天津	.2327827	.0430095	-.1851475	1.904427	2	9	4.4272435	2	1
3	河北	-.5192236	-.0215061	-.4197636	-.0583391	3	2	1.556306	2	1
4	山西	-.2078589	-.8926602	-.2751256	-.5914798	4	10	2.442894	2	1
5	内蒙古	-.3863757	.6143994	-.0850018	-.6589216	5	3	.34396065	2	1
6	辽宁	-.6192325	-.4909025	-.8162181	.552722	6	7	.24177225	2	1
7	吉林	-.7004963	.1837896	-.7007831	-.1394611	7	16	.49672512	2	1
8	黑龙江	-.8764872	.6175225	-.7598766	-.2446946	8	12	.65157546	2	1
9	上海	2.744534	.4372417	.9158435	2.528599	9	14	.44309643	1	1
10	江苏	.3243138	1.123207	.5200769	1.038666	10	17	.56872284	2	1
11	浙江	1.782139	3.779561	2.909354	.8466673	11	29	.93349776	4	2
12	安徽	-.4867221	-.0903049	-.3652568	-.5288357	12	6	.99642891	2	1
13	福建	.4865569	.5600563	2.106966	.7204288	13	28	.52560714	2	1
14	江西	-.5702236	-.3525019	-.6970371	-.3460047	14	31	1.1017742	2	1
15	山东	.5832218	-.4727882	-.1567856	-.601436	15	4	.65904896	2	1
16	河南	-.6224905	.0291783	-.5855773	-.2627164	16	27	1.0187128	2	1
17	湖北	-.4244509	-.6880782	-.4589045	-.2700696	17	20	.42418311	2	1
18	湖南	-.3305052	.6580344	.177594	-.3206295	18	22	.70076941	2	1
19	广东	1.915398	1.167288	1.872044	-.0439652	19	23	1.3086976	2	1
20	广西	-.2605374	-.3902475	.0287515	-.208404	20	5	.43407789	2	1
21	海南	-.6324823	.0312902	.1970879	-.7896157	21	18	.77219715	2	1
22	重庆	.0986308	-.4718364	.121176	-.0193789	22	21	1.0480483	2	1
23	四川	-.7282998	-.4032756	.2984568	-.125363	23	8	.39210278	2	1
24	贵州	-.8607938	.4157366	-.6277761	-.5534505	24	24	.50638234	2	1
25	云南	-.5914577	-.1654393	1.954759	.2183442	25	30	1.4209586	2	1
26	西藏	1.148904	-1.808074	-1.23813	-2.248093	26	15	3.0458377	3	1
27	陕西	-.5186969	-1.403134	-.2845287	-.313894	27	13	1.4013262	2	1
28	甘肃	-.461506	-.824605	-1.015133	-.6959153	28	25	2.2993799	2	1
29	青海	-.8206413	-.4094625	-.8431275	.0066801	29	19	3.4891544	2	1
30	宁夏	-.7557332	.9089285	-.825086	-.5872713	30	26	5.5369152	2	1
31	新疆	-.4340971	-.3834955	-.9009979	-.9565053	31	11	.	2	1

图 9.51 平均联结法聚类分析 type2 数据

在图 9.51 中，可以看到所有的观测样本被分为两类，其中浙江被分到第 2 类，其他省市为第 1 类。第 2 类的特征是经营净收入、财产性收入高，第 1 类的特征不明显。处理结果与最短联结法聚类分析是一致的。

4．加权平均联结法聚类分析

图 9.52 展示的是设定聚类数为 4，然后进行分析的结果。在输入第 1 条 Stata 命令并且分别按键盘上的回车键进行确认后，选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 9.52 所示的 type1 数据。



	v1	zv2	zv3	zv4	zv5	_clus_1_id	_clus_1_ord	_clus_1_hgt	type1
1	北京	2.491832	-1.300932	.1381473	2.74791	1	1	1.9333945	1
2	天津	.2327827	.0430095	-.1851475	1.904427	2	9	3.7787763	2
3	河北	-.5192236	-.0215061	-.4197636	-.0583391	3	2	1.556306	2
4	山西	-.2078589	-.8926602	-.2751256	-.5914798	4	10	2.4456021	2
5	内蒙古	-.3863757	.6143994	-.0850018	-.6589216	5	3	.34396065	2
6	辽宁	-.6192325	-.4909025	-.8162181	.552722	6	7	.24177225	2
7	吉林	-.7004963	.1837896	-.7007831	-.1394611	7	16	.49247288	2
8	黑龙江	-.8764872	.6175225	-.7598766	-.2446946	8	12	.675385	2
9	上海	2.744534	.4372417	.9158435	2.528599	9	14	.44309643	1
10	江苏	.3243138	1.123207	.5200769	1.038666	10	17	.56872284	2
11	浙江	1.782139	3.779561	2.909354	.8466673	11	29	.91761047	3
12	安徽	-.4867221	-.0903049	-.3652568	-.5288357	12	28	.52560714	2
13	福建	.4865569	.5600563	2.106966	.7204288	13	31	1.1271187	2
14	江西	-.5702236	-.3525019	-.6970371	-.3460047	14	4	.65904896	2
15	山东	.5832218	-.4727882	-.1567856	-.601436	15	27	1.0414301	2
16	河南	-.6224905	.0291783	-.5855773	-.2627164	16	20	.42418311	2
17	湖北	-.4244509	-.6880782	-.4589045	-.2700696	17	22	.70076941	2
18	湖南	-.3305052	.6580344	.177594	-.3206295	18	23	1.2519968	2
19	广东	1.915398	1.167288	1.872044	-.0439652	19	6	1.5324006	2
20	广西	-.2605374	-.3902475	.0287515	-.208404	20	15	1.6267549	2
21	海南	-.6324823	.0312902	.1970879	-.7896157	21	5	.43407789	2
22	重庆	.0986308	-.4718364	.121176	-.0193789	22	18	.77219715	2
23	四川	-.7282998	-.4032756	.2984568	-.125363	23	21	1.1128042	2
24	贵州	-.8607938	.4157366	-.6277761	-.5534505	24	8	.39210278	2
25	云南	-.5914577	-.1654393	1.954759	.2183442	25	24	.50638234	2
26	西藏	1.148904	-1.808074	-1.23813	-2.248093	26	30	2.895477	4
27	陕西	-.5186969	-1.403134	-.2845287	-.313894	27	13	1.4013262	2
28	甘肃	-.461506	-.824605	-1.015133	-.6959153	28	25	2.2993799	2
29	青海	-.8206413	-.4094625	-.8431275	.0066801	29	19	4.7783081	2
30	宁夏	-.7557332	.9089285	-.825086	-.5872713	30	11	6.3446125	2
31	新疆	-.4340971	-.3834955	-.9009979	-.9565053	31	26	.	2

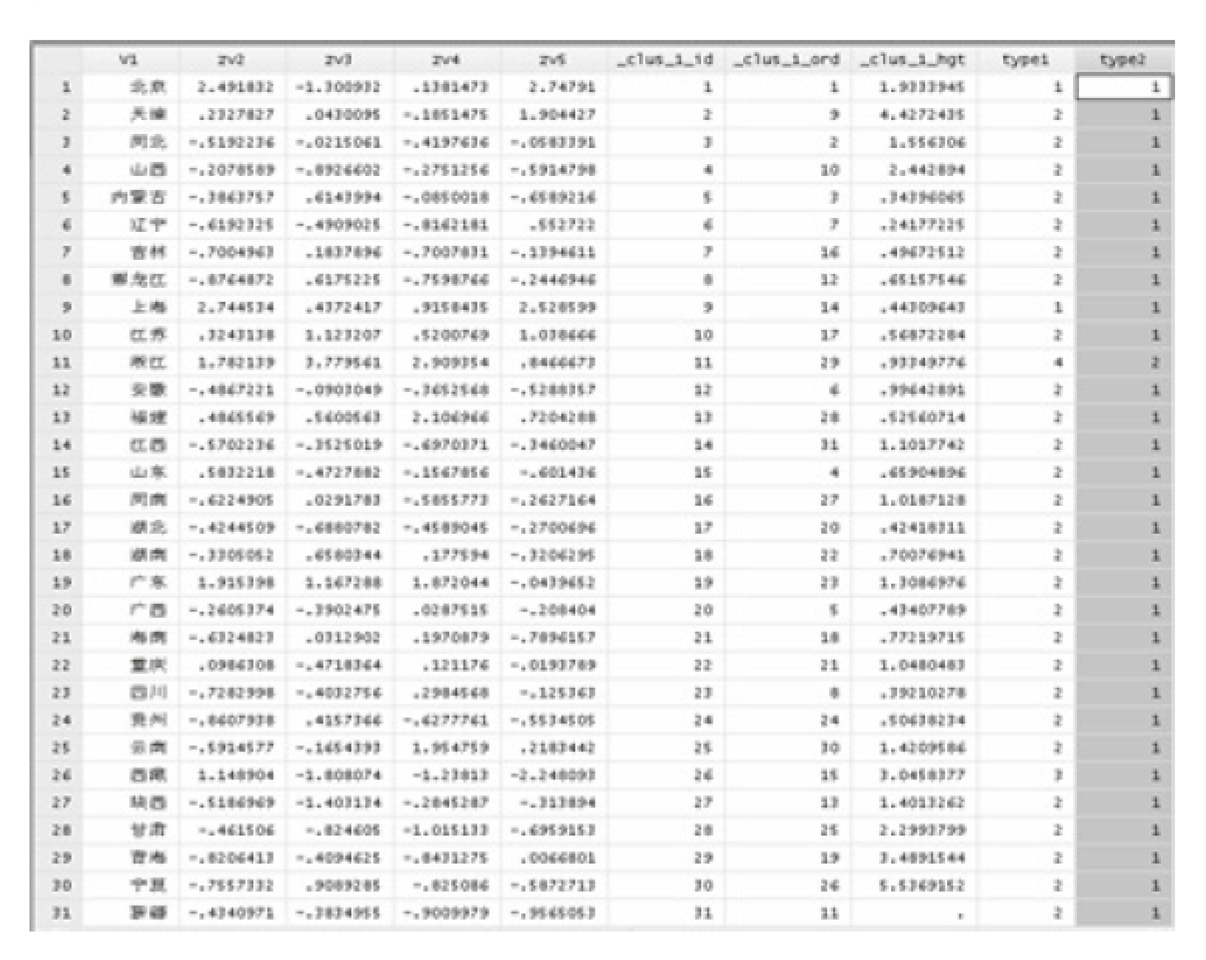
图 9.52 加权平均联结法 type1 数据

在图 9.52 中，可以看到所有的观测样本被分为 4 类：其中，浙江被分到第 3 类，上海、

北京为第 1 类，西藏为第 4 类，其他省市为第 2 类。第 3 类的特征是经营净收入、财产性收入高；第 1 类的特征是工薪收入、转移性收入高；第 4 类的特征是收入水平普遍较低；第 2 类的特征是所有收入都处在中间水平。处理结果与最短联结法聚类分析是一致的。

图 9.53 展示的是设定聚类数为 2，然后进行分析的结果。在输入第 2 条 Stata 命令并且分别按键盘上的回车键进行确认后，选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 9.53 所示的 type2 数据。

在图 9.53 中，可以看到所有的观测样本被分为两类：其中，浙江被分到第 2 类，其他省市为第 1 类。第 2 类的特征是经营净收入、财产性收入高，第 1 类的特征不明显。处理结果与最短联结法聚类分析是一致的。

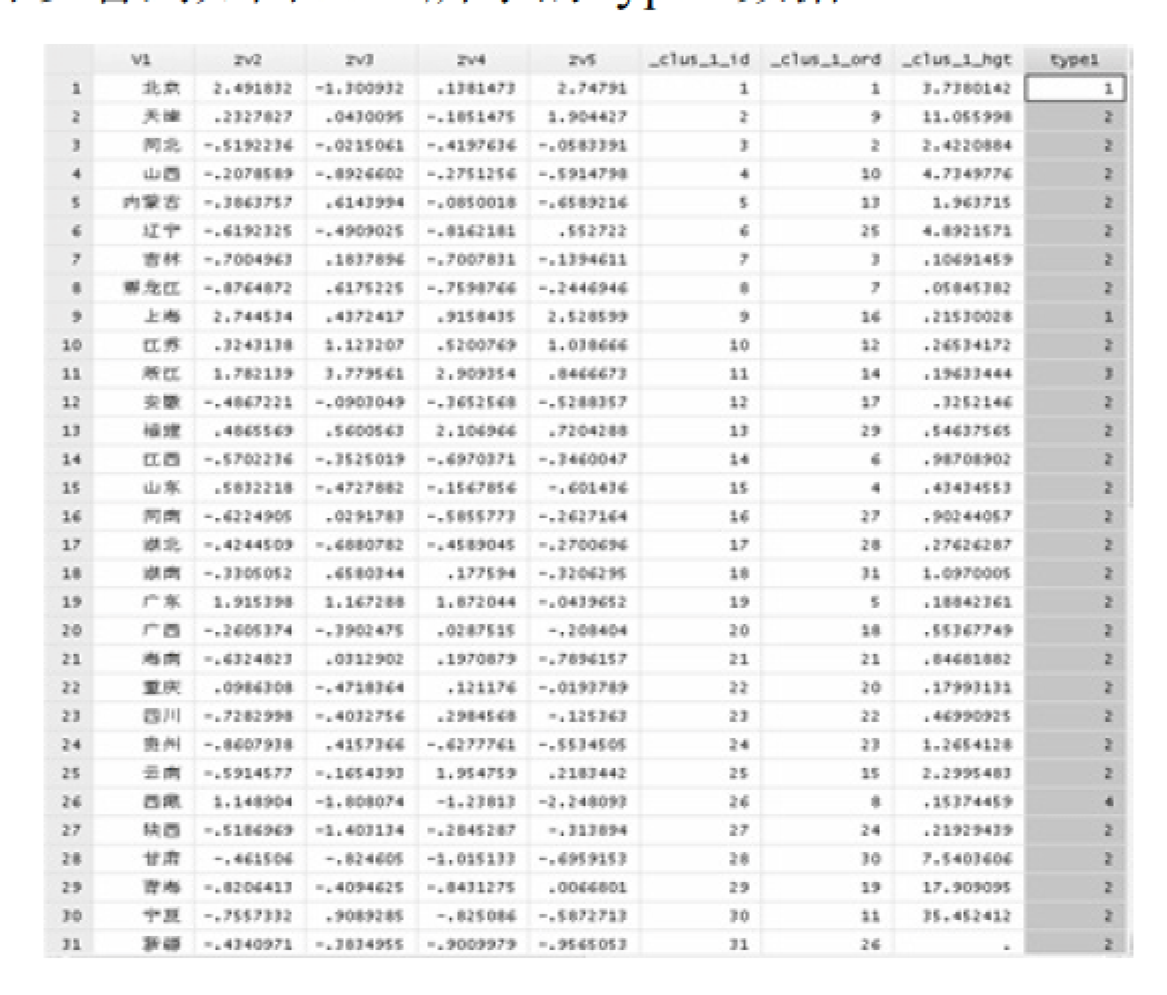


	v1	zv2	zv3	zv4	zv5	_clus_1_id	_clus_1_ord	_clus_1_hgt	type1	type2
1	北京	2.491832	-1.300932	.1381473	2.74791	1	1	1.9333945	1	1
2	天津	.2327827	.0430095	-.1851475	1.904427	2	9	4.4272435	2	1
3	河北	-.5192236	-.0215061	-.4197636	-.0583391	3	2	1.556306	2	1
4	山西	-.2078589	-.8926602	-.2751256	-.5914798	4	10	2.442894	2	1
5	内蒙古	-.3863757	.6143994	-.0850018	-.6589216	5	3	.34396065	2	1
6	辽宁	-.6192325	-.4909025	-.8162181	.552722	6	7	.24177225	2	1
7	吉林	-.7004963	.1837896	-.7007831	-.1394611	7	16	.49672512	2	1
8	黑龙江	-.8764872	.6175225	-.7598766	-.2446946	8	12	.65157546	2	1
9	上海	2.744534	.4372417	.9158435	2.528599	9	14	.44309643	1	1
10	江苏	.3243138	1.123207	.5200769	1.038666	10	17	.56872284	2	1
11	浙江	1.782139	3.779561	2.909354	.8466673	11	29	.93349776	4	2
12	安徽	-.4867221	-.0903049	-.3652568	-.5288357	12	6	.99642891	2	1
13	福建	.4865569	.5600563	2.106966	.7204288	13	28	.52560714	2	1
14	江西	-.5702236	-.3525019	-.6970371	-.3460047	14	31	1.1017742	2	1
15	山东	.5832218	-.4727882	-.1567856	-.601436	15	4	.65904896	2	1
16	河南	-.6224905	.0291783	-.5855773	-.2627164	16	27	1.0187128	2	1
17	湖北	-.4244509	-.6880782	-.4589045	-.2700696	17	20	.42418311	2	1
18	湖南	-.3305052	.6580344	.177594	-.3206295	18	22	.70076941	2	1
19	广东	1.915398	1.167288	1.872044	-.0439652	19	23	1.3086976	2	1
20	广西	-.2605374	-.3902475	.0287515	-.208404	20	5	.43407789	2	1
21	海南	-.6324823	.0312902	.1970879	-.7896157	21	18	.77219715	2	1
22	重庆	.0986308	-.4718364	.121176	-.0193789	22	21	1.0480483	2	1
23	四川	-.7282998	-.4032756	.2984568	-.125363	23	8	.39210278	2	1
24	贵州	-.8607938	.4157366	-.6277761	-.5534505	24	24	.50638234	2	1
25	云南	-.5914577	-.1654393	1.954759	.2183442	25	30	1.4209586	2	1
26	西藏	1.148904	-1.808074	-1.23813	-2.248093	26	15	3.0458377	3	1
27	陕西	-.5186969	-1.403134	-.2845287	-.313894	27	13	1.4013262	2	1
28	甘肃	-.461506	-.824605	-1.015133	-.6959153	28	25	2.2993799	2	1
29	青海	-.8206413	-.4094625	-.8431275	.0066801	29	19	3.4891544	2	1
30	宁夏	-.7557332	.9089285	-.825086	-.5872713	30	26	5.5369152	2	1
31	新疆	-.4340971	-.3834955	-.9009979	-.9565053	31	11	.	2	1

图 9.53　加权平均联结法 type2 数据

5．中位数联结法聚类分析

图 9.54 展示的是设定聚类数为 4，然后进行分析的结果。在输入第 1 条 Stata 命令并且分别按键盘上的回车键进行确认后，选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 9.54 所示的 type1 数据。



	v1	zv2	zv3	zv4	zv5	_clus_1_id	_clus_1_ord	_clus_1_hgt	type1
1	北京	2.491832	-1.300932	.1381473	2.74791	1	1	3.7380142	1
2	天津	.2327827	.0430095	-.1851475	1.904427	2	9	11.055998	2
3	河北	-.5192236	-.0215061	-.4197636	-.0583391	3	2	2.4220884	2
4	山西	-.2078589	-.8926602	-.2751256	-.5914798	4	10	4.7349776	2
5	内蒙古	-.3863757	.6143994	-.0850018	-.6589216	5	13	1.963715	2
6	辽宁	-.6192325	-.4909025	-.8162181	.552722	6	25	4.8921571	2
7	吉林	-.7004963	.1837896	-.7007831	-.1394611	7	3	.10691459	2
8	黑龙江	-.8764872	.6175225	-.7598766	-.2446946	8	7	.05845382	2
9	上海	2.744534	.4372417	.9158435	2.528599	9	16	.21530028	1
10	江苏	.3243138	1.123207	.5200769	1.038666	10	12	.26534172	2
11	浙江	1.782139	3.779561	2.909354	.8466673	11	14	.19633444	3
12	安徽	-.4867221	-.0903049	-.3652568	-.5288357	12	17	.3252146	2
13	福建	.4865569	.5600563	2.106966	.7204288	13	29	.54637565	2
14	江西	-.5702236	-.3525019	-.6970371	-.3460047	14	6	.98708902	2
15	山东	.5832218	-.4727882	-.1567856	-.601436	15	4	.43434553	2
16	河南	-.6224905	.0291783	-.5855773	-.2627164	16	27	.90244057	2
17	湖北	-.4244509	-.6880782	-.4589045	-.2700696	17	28	.27626287	2
18	湖南	-.3305052	.6580344	.177594	-.3206295	18	31	1.0970005	2
19	广东	1.915398	1.167288	1.872044	-.0439652	19	5	.18842361	2
20	广西	-.2605374	-.3902475	.0287515	-.208404	20	18	.55367749	2
21	海南	-.6324823	.0312902	.1970879	-.7896157	21	21	.84681882	2
22	重庆	.0986308	-.4718364	.121176	-.0193789	22	20	.17993131	2
23	四川	-.7282998	-.4032756	.2984568	-.125363	23	22	.46990925	2
24	贵州	-.8607938	.4157366	-.6277761	-.5534505	24	23	1.2654128	2
25	云南	-.5914577	-.1654393	1.954759	.2183442	25	15	2.2995483	2
26	西藏	1.148904	-1.808074	-1.23813	-2.248093	26	8	.15374459	4
27	陕西	-.5186969	-1.403134	-.2845287	-.313894	27	24	.21929439	2
28	甘肃	-.461506	-.824605	-1.015133	-.6959153	28	30	7.5403606	2
29	青海	-.8206413	-.4094625	-.8431275	.0066801	29	19	17.909095	2
30	宁夏	-.7557332	.9089285	-.825086	-.5872713	30	11	35.452412	2
31	新疆	-.4340971	-.3834955	-.9009979	-.9565053	31	26	.	2

图 9.54　中位数联结法 type1 数据

在图 9.54 中，可以看到所有的观测样本被分为 4 类：其中，浙江被分到第 3 类，上海、北京为第 1 类，西藏为第 4 类，其他省市为第 2 类。第 3 类的特征是经营净收入、财产性收入高；第 1 类的特征是工薪收入、转移性收入高；第 4 类的特征是收入水平普遍较低；第 2 类的特征是所有收入都处在中间水平。处理结果与最短联结法聚类分析是一致的。

图 9.55 展示的是设定聚类数为 2，然后进行分析的结果。在输入第 2 条 Stata 命令并且分别按键盘上的回车键进行确认后，选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 9.55 所示的 type2 数据。

	v1	zv2	zv3	zv4	zv5	_clus_1_id	_clus_1_ord	_clus_1_hgt	type1	type2
1	北京	2.491832	-1.300932	.1381473	2.74791	1	1	3.7380142	1	1
2	天津	.2327827	.0430095	-.1851475	1.904427	2	9	11.055998	2	1
3	河北	-.5192236	-.0215061	-.4197636	-.0583391	3	2	2.4220884	2	1
4	山西	-.2078589	-.8926602	-.2751256	-.5914798	4	10	4.7349776	2	1
5	内蒙古	-.3863757	.6143994	-.0850018	-.6589216	5	13	1.963715	2	1
6	辽宁	-.6192325	-.4909025	-.8162181	.552722	6	25	4.8921571	2	1
7	吉林	-.7004963	.1837896	-.7007831	-.1394611	7	3	.10691459	2	1
8	黑龙江	-.8764872	.6175225	-.7598766	-.2446946	8	7	.05845382	2	1
9	上海	2.744534	.4372417	.9158435	2.528599	9	16	.21530028	1	1
10	江苏	.3243138	1.123207	.5200769	1.038666	10	12	.26534172	2	1
11	浙江	1.782139	3.779561	2.909354	.8466673	11	14	.19633444	3	1
12	安徽	-.4867221	-.0903049	-.3652568	-.5288357	12	17	.3252146	2	1
13	福建	.4865569	.5600563	2.106966	.7204288	13	29	.54637565	2	1
14	江西	-.5702236	-.3525019	-.6970371	-.3460047	14	6	.98708902	2	1
15	山东	.5832218	-.4727882	-.1567856	-.601436	15	4	.43434553	2	1
16	河南	-.6224905	.0291783	-.5855773	-.2627164	16	27	.90244057	2	1
17	湖北	-.4244509	-.6880782	-.4589045	-.2700696	17	28	.27626287	2	1
18	湖南	-.3305052	.6580344	.177594	-.3206295	18	31	1.0970005	2	1
19	广东	1.915398	1.167288	1.872044	-.0439652	19	5	.18842361	2	1
20	广西	-.2605374	-.3902475	.0287515	-.208404	20	18	.55367749	2	1
21	海南	-.6324823	.0312902	.1970879	-.7896157	21	21	.84681882	2	1
22	重庆	.0986308	-.4718364	.121176	-.0193789	22	20	.17993131	2	1
23	四川	-.7282998	-.4032756	.2984568	-.125363	23	22	.46990925	2	1
24	贵州	-.8607938	.4157366	-.6277761	-.5534505	24	23	1.2654128	2	1
25	云南	-.5914577	-.1654393	1.954759	.2183442	25	15	2.2995483	2	1
26	西藏	1.148904	-1.808074	-1.23813	-2.248093	26	8	.15374459	4	2
27	陕西	-.5186969	-1.403134	-.2845287	-.313894	27	24	.21929439	2	1
28	甘肃	-.461506	-.824605	-1.015133	-.6959153	28	30	7.5403606	2	1
29	青海	-.8206413	-.4094625	-.8431275	.0066801	29	19	17.909095	2	1
30	宁夏	-.7557332	.9089285	-.825086	-.5872713	30	11	35.452412	2	1
31	新疆	-.4340971	-.3834955	-.9009979	-.9565053	31	26	.	2	1

图 9.55　中位数联结法 type2 数据

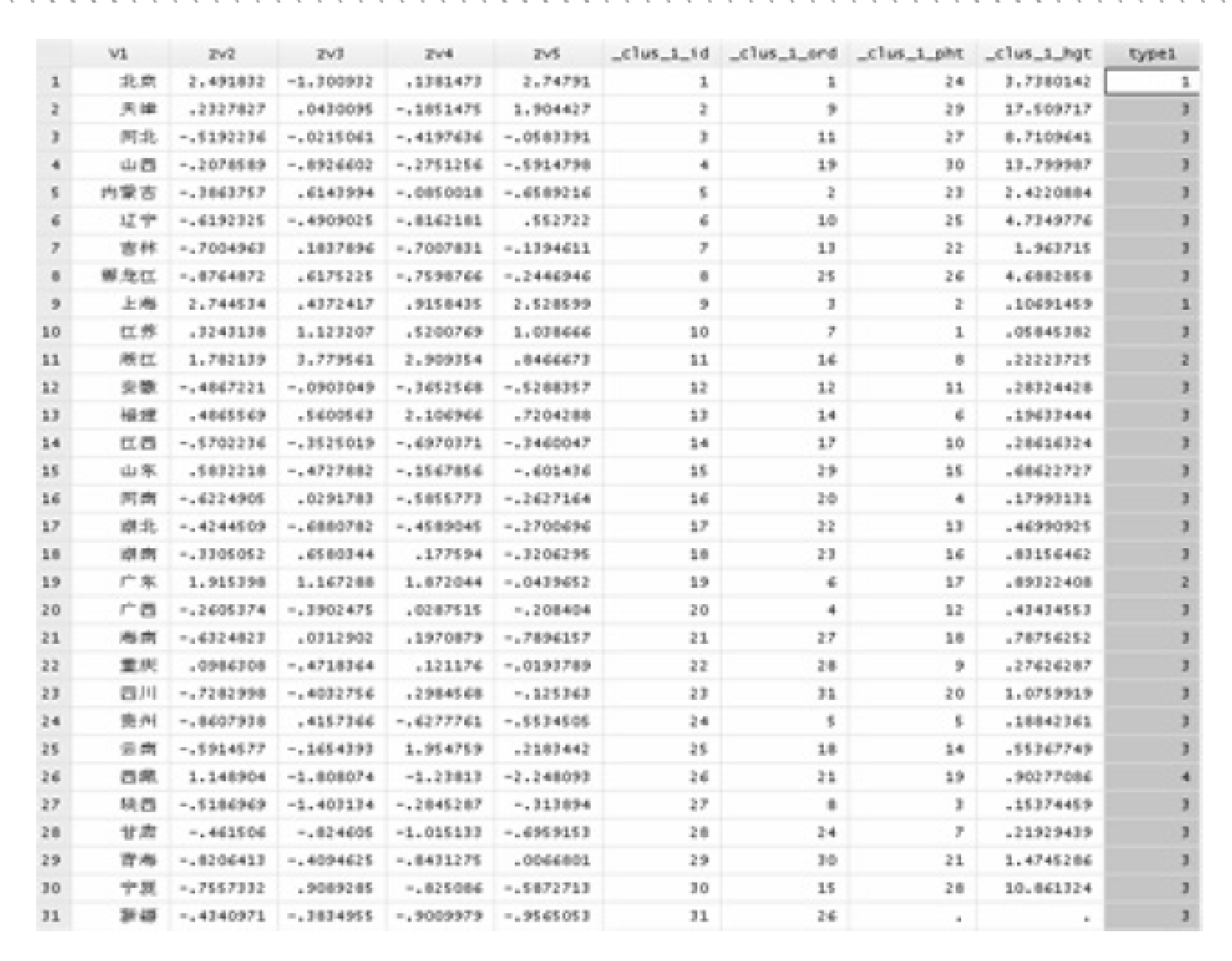
在图 9.55 中，可以看到所有的观测样本被分为两类：其中，西藏被分到第 2 类，其他省市为第 1 类。第 2 类的特征是工薪收入较高，经营净收入、财产性收入高、转移性收入较低。

6．重心联结法聚类分析

图 9.56 展示的是设定聚类数为 4，然后进行分析的结果。在输入第 7 条 Stata 命令并且分别按键盘上的回车键进行确认后，选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 9.56 所示的 type1 数据。

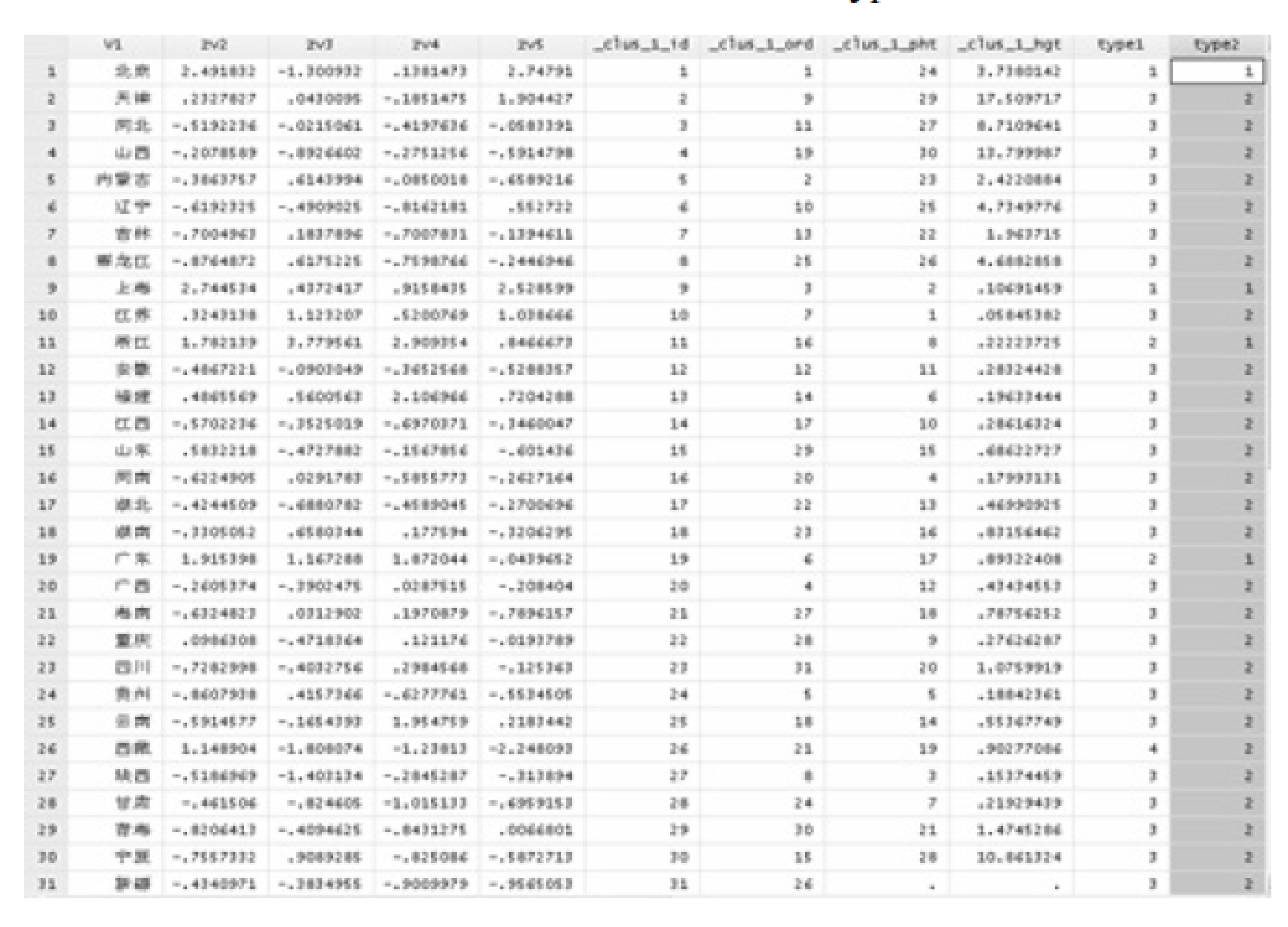
在图 9.56 中，可以看到所有的观测样本被分为 4 类：其中，浙江被分到第 2 类，上海、北京为第 1 类，西藏为第 4 类，其他省市为第 3 类。第 2 类的特征是经营净收入、财产性收入高；第 1 类的特征是工薪收入、转移性收入高；第 4 类的特征是收入水平普遍较低；第 3 类的特征是所有收入都处在中间水平。处理结果与最短联结法聚类分析是一致的。

图 9.57 展示的是设定聚类数为 2，然后进行分析的结果。在输入第 1 条 Stata 命令并且分别按键盘上的回车键进行确认后，选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 9.57 所示的 type2 数据。



	V1	zv2	zv3	zv4	zv5	_clus_1_id	_clus_1_ord	_clus_1_pht	_clus_1_hgt	type1
1	北京	2.491832	-1.300932	.1381473	2.74791	1	1	24	3.7380142	1
2	天津	.2327827	.0430095	-.1851475	1.904427	2	9	29	17.509717	3
3	河北	-.5192236	-.0215061	-.4197636	-.0583391	3	11	27	8.7109641	3
4	山西	-.2078589	-.8926602	-.2751256	-.5914798	4	19	30	13.799987	3
5	内蒙古	-.3863757	.6143994	-.0850018	-.6589216	5	2	23	2.4220884	3
6	辽宁	-.6192325	-.4909025	-.8162181	.552722	6	10	25	4.7349776	3
7	吉林	-.7004963	.1837896	-.7007831	-.1394611	7	13	22	1.963715	3
8	黑龙江	-.8764872	.6175225	-.7598766	-.2446946	8	25	26	4.6882858	3
9	上海	2.744534	.4372417	.9158435	2.528599	9	3	2	.10691459	1
10	江苏	.3243138	1.123207	.5200769	1.038666	10	7	1	.05845382	3
11	浙江	1.782139	3.779561	2.909354	.8466673	11	16	8	.22223725	2
12	安徽	-.4867221	-.0903049	-.3652568	-.5288357	12	12	11	.28324428	3
13	福建	.4865569	.5600563	2.106966	.7204288	13	14	6	.19633444	3
14	江西	-.5702236	-.3525019	-.6970371	-.3460047	14	17	10	.28616324	3
15	山东	.5832218	-.4727882	-.1567856	-.601436	15	29	15	.68622727	3
16	河南	-.6224905	.0291783	-.5855773	-.2627164	16	20	4	.17993131	3
17	湖北	-.4244509	-.6880782	-.4589045	-.2700696	17	22	13	.46990925	3
18	湖南	-.3305052	.6580344	.177594	-.3206295	18	23	16	.83156462	3
19	广东	1.915398	1.167288	1.872044	-.0439652	19	6	17	.89322408	2
20	广西	-.2605374	-.3902475	.0287515	-.208404	20	4	12	.43434553	3
21	海南	-.6324823	.0312902	.1970879	-.7896157	21	27	18	.78756252	3
22	重庆	.0986308	-.4718364	.121176	-.0193789	22	28	9	.27626287	3
23	四川	-.7282998	-.4032756	.2984568	-.125363	23	31	20	1.0759919	3
24	贵州	-.8607938	.4157366	-.6277761	-.5534505	24	5	5	.18842361	3
25	云南	-.5914577	-.1654393	1.954759	.2183442	25	18	14	.55367749	3
26	西藏	1.148904	-1.808074	-1.23813	-2.248093	26	21	19	.90277086	4
27	陕西	-.5186969	-1.403134	-.2845287	-.313894	27	8	3	.15374459	3
28	甘肃	-.461506	-.824605	-1.015133	-.6959153	28	24	7	.21929439	3
29	青海	-.8206413	-.4094625	-.8431275	.0066801	29	30	21	1.4745286	3
30	宁夏	-.7557332	.9089285	-.825086	-.5872713	30	15	28	10.861324	3
31	新疆	-.4340971	-.3834955	-.9009979	-.9565053	31	26	.	.	3

图 9.56　重心联结法聚类分析 type1 数据



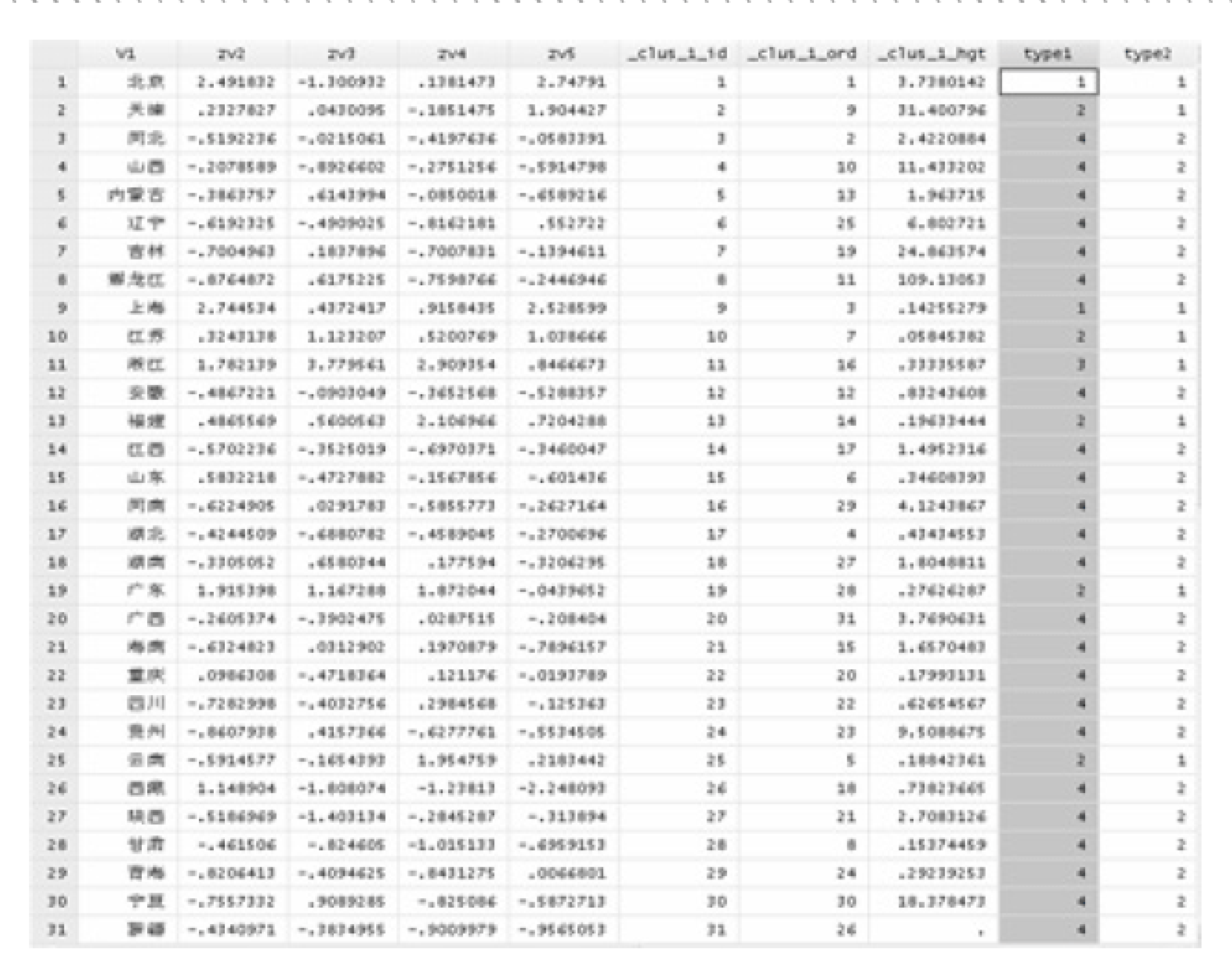
	V1	zv2	zv3	zv4	zv5	_clus_1_id	_clus_1_ord	_clus_1_pht	_clus_1_hgt	type1	type2
1	北京	2.491832	-1.300932	.1381473	2.74791	1	1	24	3.7380142	1	1
2	天津	.2327827	.0430095	-.1851475	1.904427	2	9	29	17.509717	3	2
3	河北	-.5192236	-.0215061	-.4197636	-.0583391	3	11	27	8.7109641	3	2
4	山西	-.2078589	-.8926602	-.2751256	-.5914798	4	19	30	13.799987	3	2
5	内蒙古	-.3863757	.6143994	-.0850018	-.6589216	5	2	23	2.4220884	3	2
6	辽宁	-.6192325	-.4909025	-.8162181	.552722	6	10	25	4.7349776	3	2
7	吉林	-.7004963	.1837896	-.7007831	-.1394611	7	13	22	1.963715	3	2
8	黑龙江	-.8764872	.6175225	-.7598766	-.2446946	8	25	26	4.6882858	3	2
9	上海	2.744534	.4372417	.9158435	2.528599	9	3	2	.10691459	1	1
10	江苏	.3243138	1.123207	.5200769	1.038666	10	7	1	.05845382	3	2
11	浙江	1.782139	3.779561	2.909354	.8466673	11	16	8	.22223725	2	1
12	安徽	-.4867221	-.0903049	-.3652568	-.5288357	12	12	11	.28324428	3	2
13	福建	.4865569	.5600563	2.106966	.7204288	13	14	6	.19633444	3	2
14	江西	-.5702236	-.3525019	-.6970371	-.3460047	14	17	10	.28616324	3	2
15	山东	.5832218	-.4727882	-.1567856	-.601436	15	29	15	.68622727	3	2
16	河南	-.6224905	.0291783	-.5855773	-.2627164	16	20	4	.17993131	3	2
17	湖北	-.4244509	-.6880782	-.4589045	-.2700696	17	22	13	.46990925	3	2
18	湖南	-.3305052	.6580344	.177594	-.3206295	18	23	16	.83156462	3	2
19	广东	1.915398	1.167288	1.872044	-.0439652	19	6	17	.89322408	2	1
20	广西	-.2605374	-.3902475	.0287515	-.208404	20	4	12	.43434553	3	2
21	海南	-.6324823	.0312902	.1970879	-.7896157	21	27	18	.78756252	3	2
22	重庆	.0986308	-.4718364	.121176	-.0193789	22	28	9	.27626287	3	2
23	四川	-.7282998	-.4032756	.2984568	-.125363	23	31	20	1.0759919	3	2
24	贵州	-.8607938	.4157366	-.6277761	-.5534505	24	5	5	.18842361	3	2
25	云南	-.5914577	-.1654393	1.954759	.2183442	25	18	14	.55367749	3	2
26	西藏	1.148904	-1.808074	-1.23813	-2.248093	26	21	19	.90277086	4	2
27	陕西	-.5186969	-1.403134	-.2845287	-.313894	27	8	3	.15374459	3	2
28	甘肃	-.461506	-.824605	-1.015133	-.6959153	28	24	7	.21929439	3	2
29	青海	-.8206413	-.4094625	-.8431275	.0066801	29	30	21	1.4745286	3	2
30	宁夏	-.7557332	.9089285	-.825086	-.5872713	30	15	28	10.861324	3	2
31	新疆	-.4340971	-.3834955	-.9009979	-.9565053	31	26	.	.	3	2

图 9.57　重心联结法聚类分析 type2 数据

在图 9.57 中，可以看到所有的观测样本被分为两类：其中，浙江、北京、上海、广东被分到第 1 类，其他省市为第 2 类。第 1 类的特征是各类收入普遍较高，第 2 类的特征是各类收入普遍较低。

7．Ward 联结法聚类分析

图 9.58 展示的是设定聚类数为 4，然后进行分析的结果。在输入第 1 条 Stata 命令并且分别按键盘上的回车键进行确认后，选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 9.58 所示的 type1 数据。

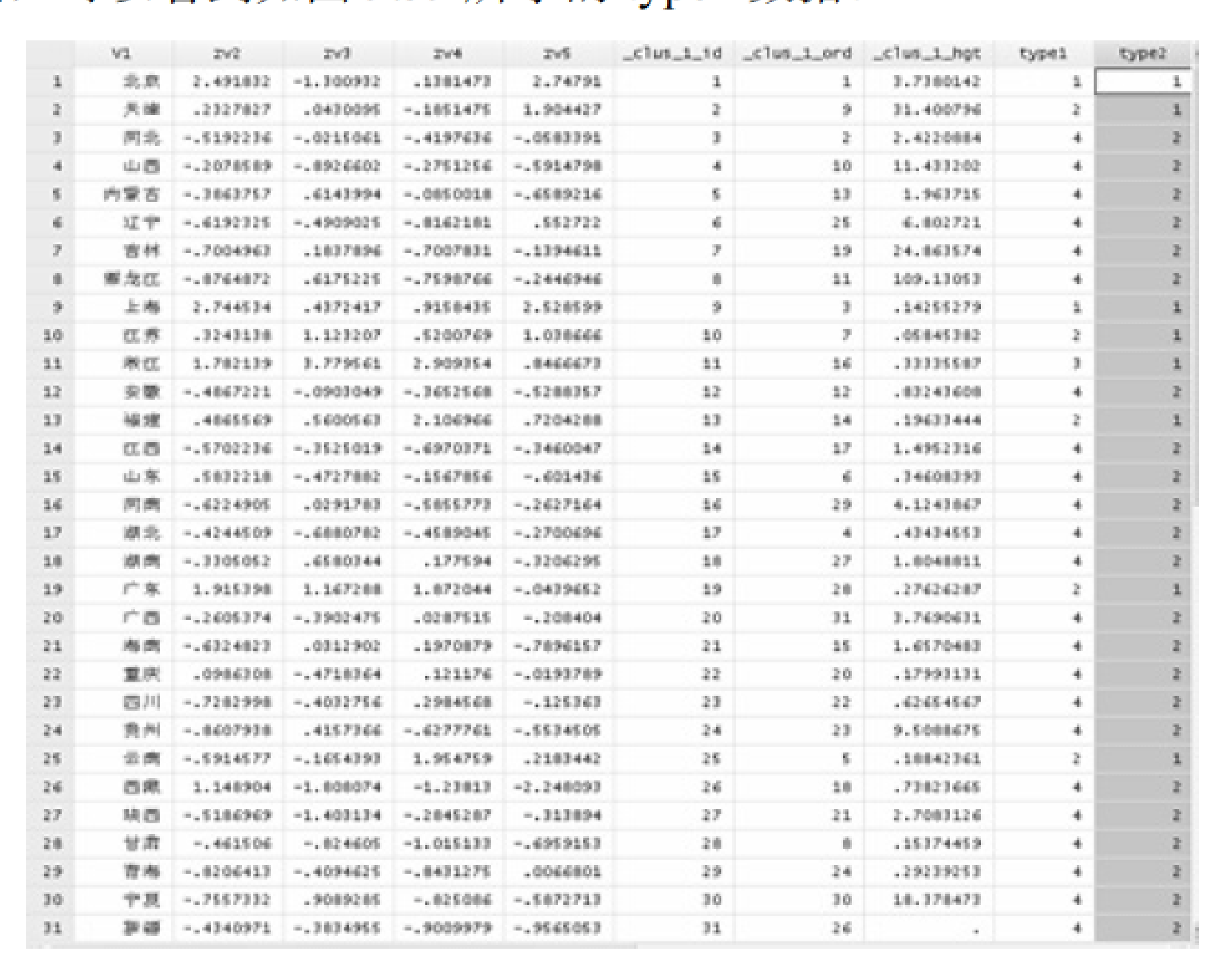


	v1	zv2	zv3	zv4	zv5	_clus_1_id	_clus_1_ord	_clus_1_hgt	type1	type2
1	北京	2.491832	-1.300932	.1381473	2.74791	1	1	3.7380142	1	1
2	天津	.2327827	.0430095	-.1851475	1.904427	2	9	31.400796	2	1
3	河北	-.5192236	-.0215061	-.4197636	-.0583391	3	2	2.4220884	4	2
4	山西	-.2078589	-.8926602	-.2751256	-.5914798	4	10	11.433202	4	2
5	内蒙古	-.3863757	.6143994	-.0850018	-.6589216	5	13	1.963715	4	2
6	辽宁	-.6192325	-.4909025	-.8162181	.552722	6	25	6.802721	4	2
7	吉林	-.7004963	.1837896	-.7007831	-.1394611	7	19	24.863574	4	2
8	黑龙江	-.8764872	.6175225	-.7598766	-.2446946	8	11	109.13053	4	2
9	上海	2.744534	.4372417	.9158435	2.528599	9	3	.14255279	1	1
10	江苏	.3243138	1.123207	.5200769	1.038666	10	7	.05845382	2	1
11	浙江	1.782139	3.779561	2.909354	.8466673	11	16	.33335587	3	1
12	安徽	-.4867221	-.0903049	-.3652568	-.5288357	12	12	.83243608	4	2
13	福建	.4865569	.5600563	2.106966	.7204288	13	14	.19633444	2	1
14	江西	-.5702236	-.3525019	-.6970371	-.3460047	14	17	1.4952316	4	2
15	山东	.5832218	-.4727882	-.1567856	-.601436	15	6	.34608393	4	2
16	河南	-.6224905	.0291783	-.5855773	-.2627164	16	29	4.1243867	4	2
17	湖北	-.4244509	-.6880782	-.4589045	-.2700696	17	4	.43434553	4	2
18	湖南	-.3305052	.6580344	.177594	-.3206295	18	27	1.8048811	4	2
19	广东	1.915398	1.167288	1.872044	-.0439652	19	28	.27626287	2	1
20	广西	-.2605374	-.3902475	.0287515	-.208404	20	31	3.7690631	4	2
21	海南	-.6324823	.0312902	.1970879	-.7896157	21	15	1.6570483	4	2
22	重庆	.0986308	-.4718364	.121176	-.0193789	22	20	.17993131	4	2
23	四川	-.7282998	-.4032756	.2984568	-.125363	23	22	.62654567	4	2
24	贵州	-.8607938	.4157366	-.6277761	-.5534505	24	23	9.5088675	4	2
25	云南	-.5914577	-.1654393	1.954759	.2183442	25	5	.18842361	2	1
26	西藏	1.148904	-1.808074	-1.23813	-2.248093	26	18	.73823665	4	2
27	陕西	-.5186969	-1.403134	-.2845287	-.313894	27	21	2.7083126	4	2
28	甘肃	-.461506	-.824605	-1.015133	-.6959153	28	8	.15374459	4	2
29	青海	-.8206413	-.4094625	-.8431275	.0066801	29	24	.29239253	4	2
30	宁夏	-.7557332	.9089285	-.825086	-.5872713	30	30	18.378473	4	2
31	新疆	-.4340971	-.3834955	-.9009979	-.9565053	31	26	.	4	2

图 9.58　Ward 联结法聚类分析 type1 数据

在图 9.58 中，可以看到所有的观测样本被分为 4 类：其中，浙江被分到第 3 类，上海、北京为第 1 类，天津、江苏、福建、广东、云南为第 2 类，其他省市为第 4 类。第 3 类的特征是经营净收入、财产性收入高；第 1 类的特征是工薪收入、转移性收入高；第 2 类的特征是收入水平普遍较高；第 4 类的特征是收入水平普遍偏低。

图 9.59 展示的是设定聚类数为 2，然后进行分析的结果。在输入第 2 条 Stata 命令并且分别按键盘上的回车键进行确认后，选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 9.59 所示的 type2 数据。



	v1	zv2	zv3	zv4	zv5	_clus_1_id	_clus_1_ord	_clus_1_hgt	type1	type2
1	北京	2.491832	-1.300932	.1381473	2.74791	1	1	3.7380142	1	1
2	天津	.2327827	.0430095	-.1851475	1.904427	2	9	31.400796	2	1
3	河北	-.5192236	-.0215061	-.4197636	-.0583391	3	2	2.4220884	4	2
4	山西	-.2078589	-.8926602	-.2751256	-.5914798	4	10	11.433202	4	2
5	内蒙古	-.3863757	.6143994	-.0850018	-.6589216	5	13	1.963715	4	2
6	辽宁	-.6192325	-.4909025	-.8162181	.552722	6	25	6.802721	4	2
7	吉林	-.7004963	.1837896	-.7007831	-.1394611	7	19	24.863574	4	2
8	黑龙江	-.8764872	.6175225	-.7598766	-.2446946	8	11	109.13053	4	2
9	上海	2.744534	.4372417	.9158435	2.528599	9	3	.14255279	1	1
10	江苏	.3243138	1.123207	.5200769	1.038666	10	7	.05845382	2	1
11	浙江	1.782139	3.779561	2.909354	.8466673	11	16	.33335587	3	1
12	安徽	-.4867221	-.0903049	-.3652568	-.5288357	12	12	.83243608	4	2
13	福建	.4865569	.5600563	2.106966	.7204288	13	14	.19633444	2	1
14	江西	-.5702236	-.3525019	-.6970371	-.3460047	14	17	1.4952316	4	2
15	山东	.5832218	-.4727882	-.1567856	-.601436	15	6	.34608393	4	2
16	河南	-.6224905	.0291783	-.5855773	-.2627164	16	29	4.1243867	4	2
17	湖北	-.4244509	-.6880782	-.4589045	-.2700696	17	4	.43434553	4	2
18	湖南	-.3305052	.6580344	.177594	-.3206295	18	27	1.8048811	4	2
19	广东	1.915398	1.167288	1.872044	-.0439652	19	28	.27626287	2	1
20	广西	-.2605374	-.3902475	.0287515	-.208404	20	31	3.7690631	4	2
21	海南	-.6324823	.0312902	.1970879	-.7896157	21	15	1.6570483	4	2
22	重庆	.0986308	-.4718364	.121176	-.0193789	22	20	.17993131	4	2
23	四川	-.7282998	-.4032756	.2984568	-.125363	23	22	.62654567	4	2
24	贵州	-.8607938	.4157366	-.6277761	-.5534505	24	23	9.5088675	4	2
25	云南	-.5914577	-.1654393	1.954759	.2183442	25	5	.18842361	2	1
26	西藏	1.148904	-1.808074	-1.23813	-2.248093	26	18	.73823665	4	2
27	陕西	-.5186969	-1.403134	-.2845287	-.313894	27	21	2.7083126	4	2
28	甘肃	-.461506	-.824605	-1.015133	-.6959153	28	8	.15374459	4	2
29	青海	-.8206413	-.4094625	-.8431275	.0066801	29	24	.29239253	4	2
30	宁夏	-.7557332	.9089285	-.825086	-.5872713	30	30	18.378473	4	2
31	新疆	-.4340971	-.3834955	-.9009979	-.9565053	31	26	.	4	2

图 9.59　Ward 联结法聚类分析 type2 数据

在图 9.59 中，可以看到所有的观测样本被分为两类：其中，浙江、北京、天津、上海、江苏、福建、广东、云南被分到第 1 类，其他省市为第 2 类。第 1 类的特征是各类收入普遍较高，第 2 类的特征是各类收入普遍较低。

9.3 本章习题

（1）表 9.4 是美国 22 家公共团体的数据。其中，1 代表该团体使用了核能源，0 代表没有使用。试利用划分聚类分析方法观测这两类企业所属类别的情况。

表 9.4 美国 22 家公共团体统计表

编号	公司	固定支出综合率/%	资产收益率/%	每千瓦容量成本/美元	每年使用的能源/万千瓦时	是否使用核能源
1	亚利桑那公共服务公司	1.06	9.2	351	9077	0
2	波士顿爱迪生公司	0.89	16.3	202	5088	1
…	…	…	…	…	…	…
21	联合装饰公司	1.04	8.4	442	6650	0
22	维吉尼亚电力公司	0.36	16.3	184	1093	1

（2）表 9.5 是我国 2006 年各地区的能源消耗情况。试用层次聚类分析方法了解我国不同地区的能源消耗情况。

表 9.5 2006 年各地区能源消耗统计表

地区	单位地区生产总值煤消耗量/吨	单位地区生产总值电消耗量/千瓦时	单位工业增加值煤消耗量/吨
北京	0.8	828.5	1.5
天津	1.11	1040.8	1.45
河北	1.96	1487.6	4.41
山西	2.95	2264.2	6.57
内蒙古	2.48	1714.1	5.67
…	…	…	…
青海	3.07	3801.8	3.44
宁夏	4.14	4997.7	9.03
新疆	2.11	1190.9	3.00

第 10 章　Stata 最小二乘线性回归分析

回归分析是经典的数据分析方法之一，应用范围非常广泛，深受学者们的喜爱。它是研究分析某一变量受到其他变量影响的分析方法，的基本思想是以被影响变量为因变量，以影响变量为自变量，研究因变量与自变量之间的因果关系。本章主要介绍最简单也最常用的最小二乘线性回归分析方法（包括简单线性回归、多重线性回归等）在具体实例中的应用。

10.1　实例一——简单线性回归分析

10.1.1　简单线性回归分析的功能与意义

Stata 的简单线性回归分析也称一元线性回归分析，是最简单也是最基本的一种回归分析方法。简单线性回归分析的特色是只涉及一个自变量，主要用来处理一个因变量与一个自变量之间的线性关系，建立变量之间的线性模型并根据模型进行评价和预测。

10.1.2　相关数据来源

	下载资源:\video\chap10\···
	下载资源:\sample\chap10\案例10.1.dta

【例 10.1】菲利普斯曲线表明，失业率和通货膨胀率之间存在着替代关系。表 10.1 给出了我国 1998－2007 年的通货膨胀率和城镇登记失业率。试用简单回归分析方法研究这种替代关系在我国是否存在。

表 10.1　我国 1998－2007 年的通货膨胀率和城镇登记失业率（单位：%）

年份	通货膨胀率	失业率
1998	−0.84	3.1
1999	−1.41	3.1
2000	0.26	3.1
2001	0.46	3.6
2002	−0.77	4.0
2003	1.16	4.3
2004	3.89	4.2
2005	1.82	4.2
2006	1.46	4.1
2007	4.75	4.0

10.1.3 Stata 分析过程

在利用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有 3 个变量，分别为年份、通货膨胀率、失业率。我们把年份变量设定为 year，把通货膨胀率变量设定为 inflation，把失业率变量设定为 unwork，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作我们在第 1 章中已有详细讲述。录入完成后数据如图 10.1 所示。

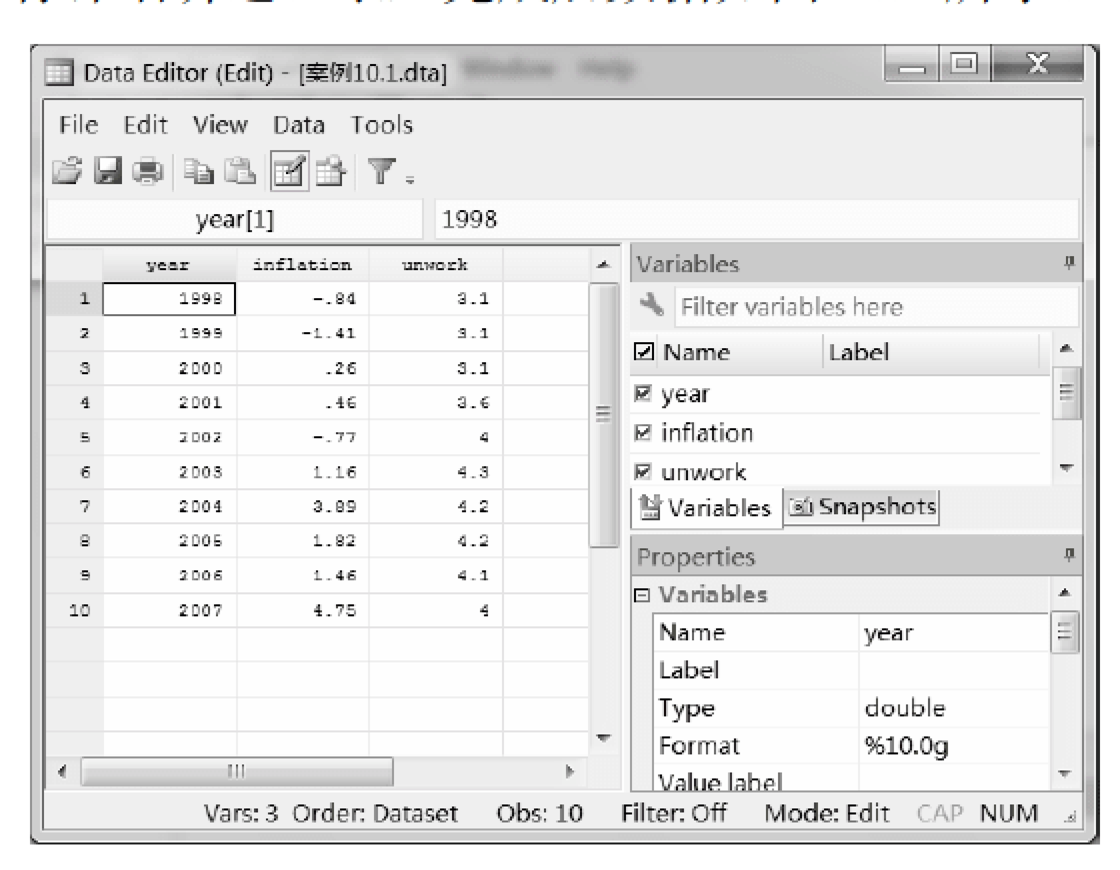


图 10.1 案例 10.1 数据

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入如下命令。

- summarize year inflation unwork,detail：本命令的含义是对年份、通货膨胀率、失业率变量进行详细描述性分析。
- correlate year inflation unwork：本命令的含义是对年份、通货膨胀率、失业率变量进行相关性分析。
- regress unwork inflation：本命令的含义是对年份、通货膨胀率、失业率变量进行简单线性回归分析。
- vce：本命令的含义是获得参与回归的各自变量的系数以及常数项的方差-协方差矩阵。
- test inflation=0：本命令的含义是检验变量通货膨胀率的系数是否显著。
- predict yhat：本命令旨在对因变量的拟合值进行预测。
- predict e,resid：本命令旨在获得回归后的残差序列

03 设置完毕后，按键盘上的回车键，等待输出结果。

10.1.4 结果分析

在 Stata 14.0 主界面的结果窗口可以看到如图 10.2~图 10.8 所示的分析结果。

1．对数据进行描述性分析的结果

图 10.2 是对数据进行描述性分析的结果。关于这一分析已在前面的章节中详细介绍过，这里不再赘述。在回归分析中，通过本步操作可以从整体上了解数据的一般特征。本步骤的操作是非常有必要的，因为有些时候数据可能会存在某些异常值（非常大或者非常小），也有些时候各个变量间的量纲差距过大，例如某个变量是几百万，同时另一个变量是零点几，那么系统有可能会把小变量忽略掉，这些都会严重影响数据的回归分析结果。

```
. summarize  year inflation unwork,detail

                            year
-------------------------------------------------------------
      Percentiles      Smallest
 1%         1998           1998
 5%         1998           1999
10%       1998.5           2000       Obs                  10
25%         2000           2001       Sum of Wgt.          10

50%       2002.5                      Mean             2002.5
                        Largest       Std. Dev.       3.02765
75%         2005           2004
90%       2006.5           2005       Variance       9.166667
95%         2007           2006       Skewness              0
99%         2007           2007       Kurtosis       1.775758

                          inflation
-------------------------------------------------------------
      Percentiles      Smallest
 1%        -1.41          -1.41
 5%        -1.41           -.84
10%       -1.125           -.77       Obs                  10
25%         -.77            .26       Sum of Wgt.          10

50%          .81                      Mean              1.078
                        Largest       Std. Dev.      2.011886
75%         1.82           1.46
90%         4.32           1.82       Variance       4.047684
95%         4.75           3.89       Skewness        .613555
99%         4.75           4.75       Kurtosis       2.326643

                            unwork
-------------------------------------------------------------
      Percentiles      Smallest
 1%          3.1            3.1
 5%          3.1            3.1
10%          3.1            3.1       Obs                  10
25%          3.1            3.6       Sum of Wgt.          10

50%            4                      Mean               3.77
                        Largest       Std. Dev.       .498999
75%          4.2            4.1
90%         4.25            4.2       Variance           .249
95%          4.3            4.2       Skewness      -.5081105
99%          4.3            4.3       Kurtosis       1.533439
```

图 10.2　描述性分析的结果

在如图 10.2 所示的分析结果中，可以得到很多信息，包括百分位数、4 个最小值、4 个最大值、平均值、标准差、偏度、峰度等。

（1）百分位数（Percentiles）

可以看出变量 year 的第 1 个四分位数（25%）是 2000，第 2 个四分位数（50%）是 2002.5，第 3 个四分位数（75%）是 2005；变量 inflation 的第 1 个四分位数（25%）是-0.77，第 2 个四分位数（50%）是 0.81，第 3 个四分位数（75%）是 1.82；变量 unwork 的第 1 个四分位数（25%）是 3.1，第 2 个四分位数（50%）是 4，第 3 个四分位数（75%）是 4.2。

（2）4 个最小值（Smallest）

变量 year 最小的 4 个数据值分别是 1998、1999、2000、2001

变量 inflation 最小的 4 个数据值分别是-1.41、-0.84、-0.77、0.26。

变量 unwork 最小的 4 个数据值分别是 3.1、3.1、3.1、3.6。

（3）4 个最大值（Largest）

变量 year 最大的 4 个数据值分别是 2004、2005、2006、2007。

变量 inflation 最大的 4 个数据值分别是 1.46、1.82、3.89、4.75。

变量 unwork 最大的 4 个数据值分别是 4.1、4.2、4.2、4.3。

（4）平均值（Mean）和标准差（Std. Dev）

变量 year 的平均值为 2002.5，标准差是 3.02765。

变量 inflation 的平均值为 1.078，标准差是 2.011886。

变量 unwork 的平均值为 3.77，标准差是 0.498999。

（5）偏度（Skewness）和峰度（Kurtosis）

变量 year 的偏度为 0，为无偏度。

变量 inflation 的偏度为 0.613555，为正偏度但不大。

变量 unwork 的偏度为-0.5081105，为负偏度但不大。

变量 year 的峰度为 1.775758，有一个比正态分布更短的尾巴。

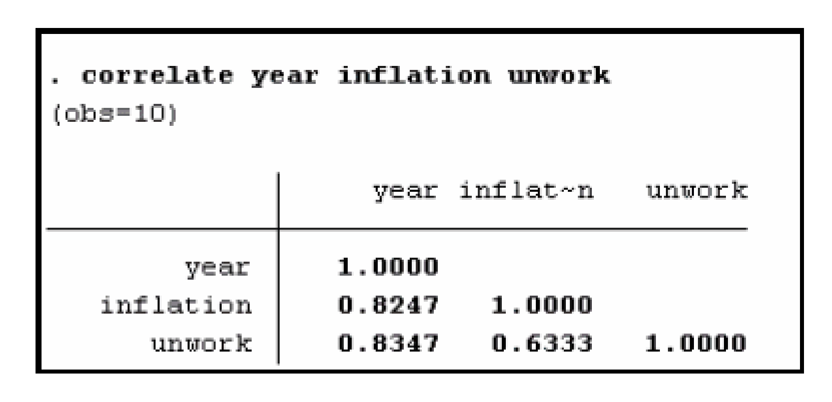
变量 inflation 的峰度为 2.326643，有一个比正态分布更短的尾巴。

变量 unwork 的峰度为 1.533439，有一个比正态分布更短的尾巴。

综上所述，数据的总体质量还是可以的，没有极端异常值，变量间的量纲差距、变量的偏度、峰度也是可以接受的，可以进入下一步的分析。

2．对数据进行相关性分析的结果

图 10.3 是对数据进行相关性分析的结果。关于这一分析我们在前面的章节中已详细介绍过，这里不再赘述。相关分析是回归分析中非常重要的一部分，因为回归分析的本意就是研究自变量对因变量的影响关系，如果参与回归分析的变量本身就是不相关的，那么回归分析就会失去意义。如果通过回归分析探索出变量之间存在着一定关系，那么这种关系也未必是真实的，它有可能仅仅是由于数据特征的某种巧合而拟合出了回归模型。综上所述，变量之间存在相关关系是进行回归分析的必要前提。



```
. correlate year inflation unwork
(obs=10)

             |     year inflat~n   unwork
-------------+---------------------------
        year |   1.0000
   inflation |   0.8247   1.0000
      unwork |   0.8347   0.6333   1.0000
```

图 10.3　相关性分析的结果

在图 10.3 中，变量通货膨胀率和失业率之间的相关系数是 0.6333，这说明两个变量之间存在较强的正相关关系，所以我们可以进行回归分析。

3．对数据进行回归分析的结果

图 10.4 是对数据进行回归分析的结果。

```
. regress unwork inflation

      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  1,     8) =    5.36
       Model |  .898891486     1  .898891486           Prob > F      =  0.0493
    Residual |  1.34210851     8  .167763564           R-squared     =  0.4011
-------------+------------------------------           Adj R-squared =  0.3263
       Total |       2.241     9        .249           Root MSE      =  .40959

------------------------------------------------------------------------------
      unwork |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   inflation |    .157083   .0678616     2.31   0.049     .0005938    .3135721
       _cons |   3.600665   .1487548    24.21   0.000     3.257635    3.943694
------------------------------------------------------------------------------
```

图 10.4　回归分析的结果

从上述分析结果中可以得到很多信息。可以看出共有 10 个样本参与了分析，模型的 F 值(1，8) ＝ 5.36，P 值（Prob ＞ F）＝ 0.0493，说明模型整体上是非常显著的。模型的可决系数（R-squared）为 0.4011，模型修正的可决系数（Adj R-squared）= 0.3263，说明模型的解释能力还是差强人意的。

模型的回归方程是：

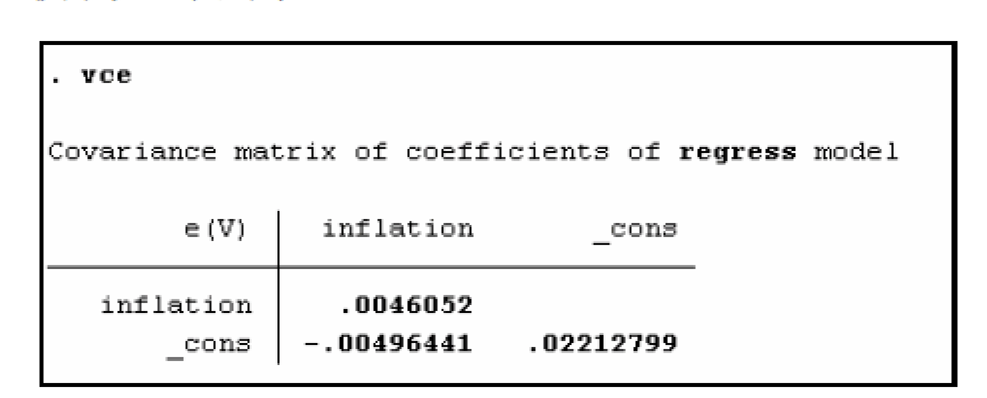
unwork=0.157083*inflation+3.600665

变量 inflation 的系数标准误是 0.0678616，t 值为 2.31，P 值为 0.049，系数是非常显著的，95%的置信区间为[0.0005938,0.3135721]。常数项的系数标准误是 0.1487548，t 值为 24.21，P 值为 0.000，系数也是非常显著的，95%的置信区间为[3.257635,3.943694]。

从上面的分析可以看出通货膨胀率和失业率之间是一种正向联动变化关系，通货膨胀率每增加一点，失业率就增加 0.157 点。通货膨胀和失业的替代关系在我国并不存在。

4．变量的方差-协方差矩阵

图 10.5 是变量的方差-协方差矩阵。



```
. vce

Covariance matrix of coefficients of regress model

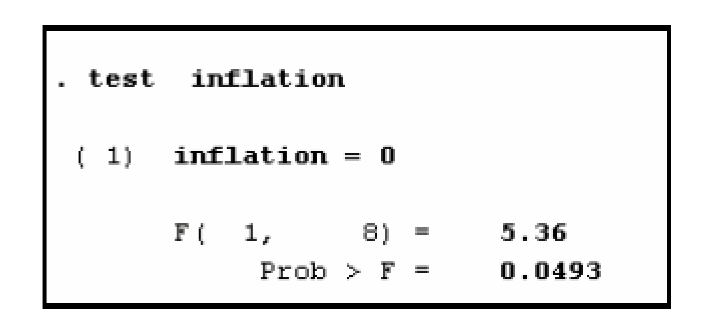
        e(V) |  inflation       _cons
-------------+------------------------
   inflation |   .0046052
       _cons | -.00496441    .02212799
```

图 10.5　变量的方差-协方差矩阵

从图 10.5 中可以看出，变量的方差与协方差都不是很大。

5．对变量系数的假设检验结果

图 10.6 是对变量系数的假设检验结果。



```
. test  inflation

 ( 1)  inflation = 0

       F(  1,     8) =    5.36
            Prob > F =    0.0493
```

图 10.6　对变量系数的假设检验结果

从图 10.6 中可以看出，通货膨胀率的系数非常显著，在 5%的显著性水平上通过了检验。

6．对因变量的拟合值的预测

图 10.7 是对因变量的拟合值的预测。

	year	inflation	unwork	yhat
1	1998	-.84	3.1	3.468715
2	1999	-1.41	3.1	3.379178
3	2000	.26	3.1	3.641506
4	2001	.46	3.6	3.672923
5	2002	-.77	4	3.479711
6	2003	1.16	4.3	3.782881
7	2004	3.89	4.2	4.211717
8	2005	1.82	4.2	3.886555
9	2006	1.46	4.1	3.830006
10	2007	4.75	4	4.346808

图 10.7　对因变量的拟合值的预测

因变量预测拟合值是根据自变量的值和得到的回归方程计算出来的，主要用于预测未来。在图 10.7 中，可以看到 yhat 的值与 unwork 的值是比较相近的，所以拟合的回归模型还是不错的。关于预测未来的作用将在案例延伸部分进行详细说明。

7．回归分析得到的残差序列

图 10.8 是回归分析得到的残差序列。

	year	inflation	unwork	yhat	e
1	1998	-.84	3.1	3.468715	-.3687149
2	1999	-1.41	3.1	3.379178	-.2791776
3	2000	.26	3.1	3.641506	-.5415062
4	2001	.46	3.6	3.672923	-.0729227
5	2002	-.77	4	3.479711	.5202893
6	2003	1.16	4.3	3.782881	.5171192
7	2004	3.89	4.2	4.211717	-.0117173
8	2005	1.82	4.2	3.886555	.3134444
9	2006	1.46	4.1	3.830006	.2699943
10	2007	4.75	4	4.346808	-.3468086

图 10.8　残差序列

残差序列是很有用处的。例如，它可以用来检验变量是否存在异方差，也可以用来检验变量间是否存在协整关系等。在后续章节中将会进行详细说明，这里不再赘述。

10.1.5　案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

1．延伸 1：在回归方程中不包含常数项

以本例为例进行说明，回归分析操作命令可以相应地修改为：

```
regress unwork inflation，nocon
```

在命令窗口输入命令并按回车键进行确认，结果如图 10.9 所示。

```
. regress unwork inflation,nocon

      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  1,     9) =    4.04
       Model |  44.7332293     1  44.7332293           Prob > F      =  0.0753
    Residual |  99.6347707     9  11.0705301           R-squared     =  0.3099
-------------+------------------------------           Adj R-squared =  0.2332
       Total |      144.37    10      14.437           Root MSE      =  3.3272

------------------------------------------------------------------------------
      unwork |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   inflation |   .9648907   .4799959     2.01   0.075    -.1209354    2.050717
------------------------------------------------------------------------------
```

图 10.9　延伸 1 分析结果图

从上述分析结果中，模型的 F 值下降为 4.04，P 值（Prob > F）上升为 0.0753，说明模型整体的显著程度有所下降。模型的可决系数（R-squared）下降为 0.3099，模型修正的可决系数（Adj R-squared）下降为 0.2332。

模型的回归方程变为：

unwork=0.9648907*inflation

变量 inflation 的系数标准误是 0.4799959，t 值为 2.01，P 值为 0.075，系数的显著程度有所下降，95%的置信区间为[-0.1209354, 2.050717]。

从上面的分析可以看出不包含常数项的回归方程不论是在模型整体的显著程度、变量系数的显著程度还是在模型的解释能力上都较包含常数项的回归方程有所下降。

2．延伸 2：限定参与回归的样本范围

以本例为例进行说明，例如我们只对 2000 年以后的样本进行回归分析，操作命令可以相应地修改为：

```
regress unwork inflation if year>=2000
```

在命令窗口输入命令并按回车键进行确认，结果如图 10.10 所示。

```
. regress unwork inflation if year>=2000

      Source |       SS       df       MS              Number of obs =       8
-------------+------------------------------           F(  1,     6) =    1.08
       Model |  .171152798     1  .171152798           Prob > F      =  0.3380
    Residual |  .947597202     6  .157932867           R-squared     =  0.1530
-------------+------------------------------           Adj R-squared =  0.0118
       Total |     1.11875     7  .159821429           Root MSE      =  .39741

------------------------------------------------------------------------------
      unwork |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   inflation |   .0842132   .0808955     1.04   0.338     -.113731    .2821574
       _cons |   3.800338   .1926186    19.73   0.000     3.329017    4.271659
------------------------------------------------------------------------------
```

图 10.10　延伸 2 分析结果图

关于结果的分析与前面类似，限于篇幅，这里不再赘述。

3．延伸 3：关于回归预测

以本例为例进行说明，例如将年份扩展至 2007 年，假定该年的通货膨胀率为 5%，把样本数据输入到数据文件中，然后进行预测，操作命令如下：

```
predict yyhat
```

在命令窗口输入命令并按回车键进行确认，结果如图 10.11 所示。

	year	inflation	unwork	yyhat
4	2001	.46	3.6	3.839076
5	2002	-.77	4	3.735494
6	2003	1.16	4.3	3.898025
7	2004	3.89	4.2	4.127927
8	2005	1.82	4.2	3.953606
9	2006	1.46	4.1	3.923289
10	2007	4.75	4	4.20035
11	2008	5	.	4.221404

图 10.11　描述性分析的结果

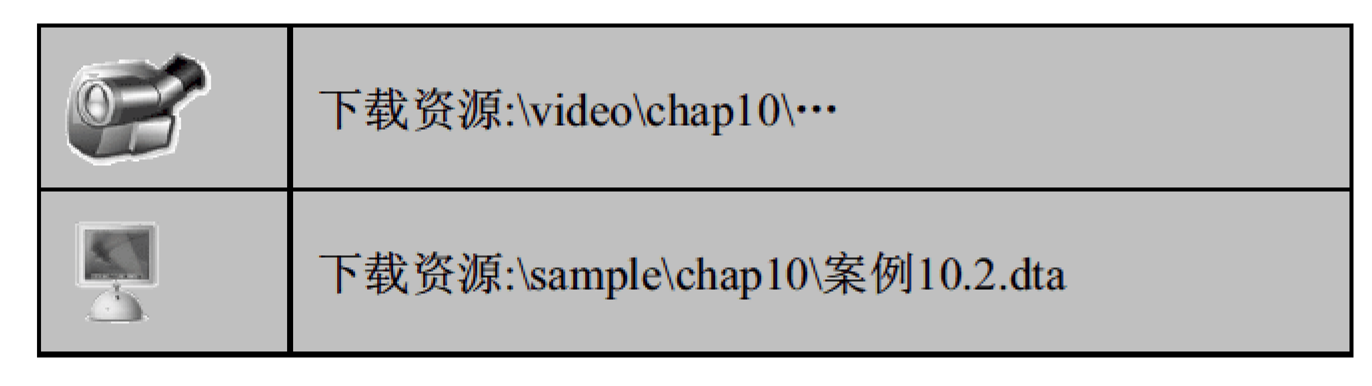
可以看到在图 10.11 中出现了预测的因变量数据，即在通货膨胀率为 5%时，预测的失业率将会是 4.221404%。

10.2　实例二——多重线性回归分析

10.2.1　多重线性回归分析的功能与意义

Stata 的多重线性回归分析也称多元线性回归分析，是最为常用的一种回归分析（Regression）方法。多重线性回归分析涉及多个自变量，用来处理一个因变量与多个自变量之间的线性关系，建立变量之间的线性模型并根据模型进行评价和预测。

10.2.2　相关数据来源



下载资源:\video\chap10\…

下载资源:\sample\chap10\案例10.2.dta

【例 10.2】为了检验美国电力行业是否存在规模经济，Nerlove（1963）收集了 1955 年 145 家美国电力企业的总成本（TC）、产量（Q）、工资率（PL）、燃料价格（PF）及资本租赁价格（PK）的数据，如表 10.2 所示。试以总成本为因变量，以产量、工资率、燃料价格和资本租赁价格为自变量，利用多重回归分析方法研究其间的关系。

表 10.2　美国电力企业相关数据

编号	TC/百万美元	Q/千瓦时	PL/美元/千瓦时	PF/美元/千瓦时	PK/美元/千瓦时
1	0.082	2	2.1	17.9	183
2	0.661	3	2.1	35.1	174
3	0.990	4	2.1	35.1	171
4	0.315	4	1.8	32.2	166
5	0.197	5	2.1	28.6	233
6	0.098	9	2.1	28.6	195
…	…	…	…	…	…
143	73.050	11796	2.1	28.6	148
144	139.422	14359	2.3	33.5	212
145	119.939	16719	2.3	23.6	162

10.2.3　Stata 分析过程

在利用 Stata 进行分析之前，要把数据录入到 Stata 中。本例中有 5 个变量，分别是总成本（TC）、产量（Q）、工资率（PL）、燃料价格（PF）及资本租赁价格（PK）。把变量类型及长度设定为系统默认方式，然后录入相关数据。相关操作在第 1 章中已有详细讲述，这里不再赘述。录入完成后数据如图 10.12 所示。

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入如下命令。

- summarize　TC Q PL PF PK,detail：本命令的含义是对总成本（TC）、产量（Q）、工资率（PL）、燃料价格（PF）及资本租赁价格（PK）变量进行详细描述性分析。
- correlate TC Q PL PF PK：本命令的含义是对总成本（TC）、产量（Q）、工资率（PL）、燃料价格（PF）及资本租赁价格（PK）变量进行相关性分析。
- regress TC Q PL PF PK：本命令的含义是对总成本（TC）、产量（Q）、工资率（PL）、燃料价格（PF）及资本租赁价格（PK）变量进行多重线性回归分析。
- vce：本命令的含义是获得参与回归的各自变量的系数以及常数项的方差-协方差矩阵。
- test Q PL PF PK：本命令的含义是检验各自变量系数的联合显著性。
- predict yhat：本命令旨在对因变量的拟合值进行预测。
- predict e,resid：本命令旨在获得回归后的残差序列。
- regress TC Q PL PF：本命令的含义是对总成本（TC）、产量（Q）、工资率（PL）、燃料价格（PF）等变量进行多重线性回归分析。

03 设置完毕后，按键盘上的回车键，等待输出结果。

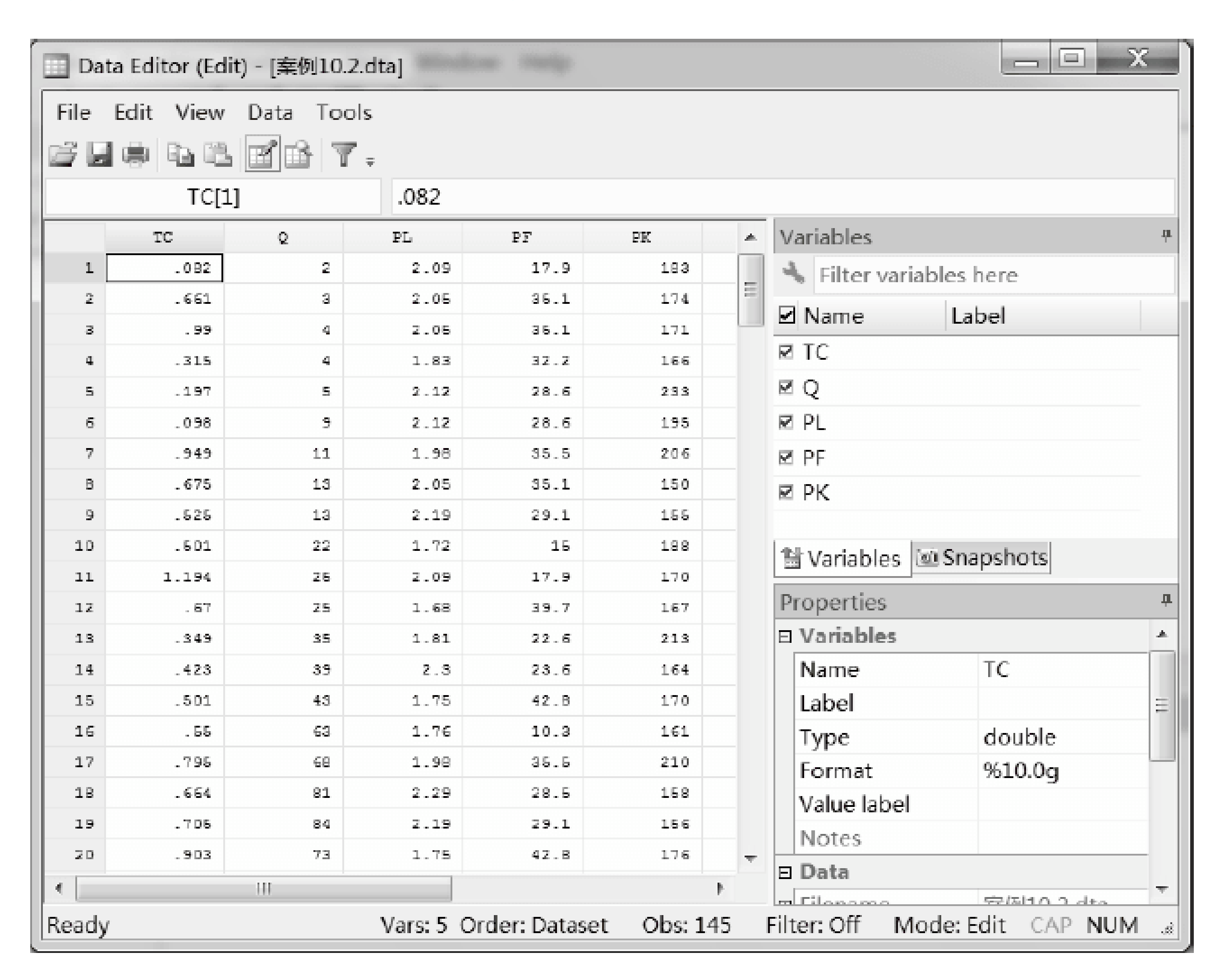


图 10.12　案例 10.2 数据

10.2.4 结果分析

在 Stata 14.0 主界面的结果窗口可以看到如图 10.13~图 10.20 所示的分析结果。

1．对数据进行描述性分析的结果

图 10.13 是对数据进行描述性分析的结果。关于这一分析过程对于回归分析的重要意义在上节已经论述过，此处不再重复讲解。

```
. summarize  TC Q PL PF PK,detail

                             TC
-------------------------------------------------------------
      Percentiles      Smallest
 1%         .098           .082
 5%         .501           .098
10%         .705           .197       Obs                 145
25%        2.382           .315       Sum of Wgt.         145

50%        6.754                      Mean            12.9761
                        Largest       Std. Dev.      19.79458
75%       14.132         69.878
90%       32.318          73.05       Variance       391.8253
95%       44.894        119.939       Skewness       3.636095
99%      119.939        139.422       Kurtosis       19.66927

                              Q
-------------------------------------------------------------
      Percentiles      Smallest
 1%            3              2
 5%           13              3
10%           43              4       Obs                 145
25%          279              4       Sum of Wgt.         145

50%         1109                      Mean           2133.083
                        Largest       Std. Dev.      2931.942
75%         2507          11477
90%         5819          11796       Variance        8596285
95%         8642          14359       Skewness       2.398202
99%        14359          16719       Kurtosis       9.474916

                             PL
-------------------------------------------------------------
      Percentiles      Smallest
 1%         1.45           1.45
 5%         1.55           1.45
10%         1.68           1.52       Obs                 145
25%         1.76           1.52       Sum of Wgt.         145

50%         2.04                      Mean           1.972069
                        Largest       Std. Dev.      .2368072
75%         2.19           2.32
90%          2.3           2.32       Variance       .0560776
95%         2.31           2.32       Skewness      -.2539563
99%         2.32           2.32       Kurtosis       1.974824

                             PF
-------------------------------------------------------------
      Percentiles      Smallest
 1%         10.3           10.3
 5%         10.3           10.3
10%         12.9           10.3       Obs                 145
25%         21.3           10.3       Sum of Wgt.         145

50%         26.9                      Mean           26.17655
                        Largest       Std. Dev.      7.876071
75%         32.2           39.7
90%         35.1           42.8       Variance        62.0325
95%         36.2           42.8       Skewness      -.3328658
99%         42.8           42.8       Kurtosis       2.641048

                             PK
-------------------------------------------------------------
      Percentiles      Smallest
 1%          143            138
 5%          155            143
10%          157            144       Obs                 145
25%          162            148       Sum of Wgt.         145

50%          170                      Mean           174.4966
                        Largest       Std. Dev.      18.20948
75%          183            225
90%          202            225       Variance       331.5851
95%          212            227       Skewness       .9992943
99%          227            233       Kurtosis       3.772226
```

图 10.13　描述性分析的结果

在如图 10.13 所示的分析结果中，可以得到很多信息，包括百分位数、4 个最小值、4 个最大值、平均值、标准差、偏度、峰度等。

（1）百分位数（Percentiles）

可以看出变量 TC 的第 1 个四分位数（25%）是 2.382，第 2 个四分位数（50%）是 6.754，第 3 个四分位数（75%）是 14.132；变量 Q 的第 1 个四分位数（25%）是 279，第 2 个四分位数（50%）是 1109，第 3 个四分位数（75%）是 2507；变量 PL 的第 1 个四分位数（25%）是 1.76，第 2 个四分位数（50%）是 2.04，第 3 个四分位数（75%）是 2.19；变量 PF 的第 1 个四分位数（25%）是 21.3，第 2 个四分位数（50%）是 26.9，第 3 个四分位数（75%）是 32.2；

变量 PK 的第 1 个四分位数（25%）是 162，第 2 个四分位数（50%）是 170，第 3 个四分位数（75%）是 183。

（2）4 个最小值（Smallest）

变量 TC 最小的 4 个数据值分别是 0.082、0.098、0.197、0.315。

变量 Q 最小的 4 个数据值分别是 2、3、4、4。

变量 PL 最小的 4 个数据值分别是 1.45、1.45、1.52、1.52。

变量 PF 最小的 4 个数据值分别是 10.3、10.3、10.3、10.3。

变量 PK 最小的 4 个数据值分别是 138、143、144、148。

（3）4 个最大值（Largest）

变量 TC 最大的 4 个数据值分别是 69.878、73.05、119.939、139.422。

变量 Q 最大的 4 个数据值分别是 11477、11796、14359、16719。

变量 PL 最大的 4 个数据值分别是 2.32、2.32、2.32、2.32。

变量 PF 最大的 4 个数据值分别是 39.7、42.8、42.8、42.8。

变量 PK 最大的 4 个数据值分别是 225、225、227、233。

（4）平均值（Mean）和标准差（Std. Dev）

变量 TC 的平均值为 12.9761，标准差是 19.79458。

变量 Q 的平均值为 2133.083，标准差是 2931.942。

变量 PL 的平均值为 1.972069，标准差是 0.2368072。

变量 PF 的平均值为 26.17655，标准差是 7.876071。

变量 PK 的平均值为 174.4966，标准差是 18.20948。

（5）偏度（Skewness）和峰度（Kurtosis）

变量 TC 的偏度为 3.636095，为正偏度但不大。

变量 Q 的偏度为 2.398202，为正偏度但不大。

变量 PL 的偏度为-0.2539563，为负偏度但不大。

变量 PF 的偏度为-0.3328658，为负偏度但不大。

变量 PK 的偏度为 0.9992943，为正偏度但不大。

变量 TC 的峰度为 19.66927，有一个比正态分布更长的尾巴。

变量 Q 的峰度为 9.474916，有一个比正态分布更长的尾巴。

变量 PL 的峰度为 1.974824，有一个比正态分布更短的尾巴。

变量 PF 的峰度为 2.641048，有一个比正态分布更短的尾巴。

变量 PK 的峰度为 3.772226，有一个比正态分布略长的尾巴。

综上所述，数据的总体质量还是可以的，没有极端异常值，变量间的量纲差距、变量的偏度、峰度也是可以接受的，可以进入下一步的分析。

2．对数据进行相关性分析的结果

图 10.14 是对数据进行相关性分析的结果。关于这一分析过程对于回归分析的重要意义在上节已经论述过，此处不再重复讲解。

```
. correlate TC Q PL PF PK
(obs=145)

             |       TC        Q       PL       PF       PK
-------------+---------------------------------------------
          TC |   1.0000
           Q |   0.9525   1.0000
          PL |   0.2513   0.1714   1.0000
          PF |   0.0339  -0.0773   0.3137   1.0000
          PK |   0.0272   0.0029  -0.1781   0.1254   1.0000
```

图 10.14　相关性分析的结果

在图 10.14 中，TC 与各个自变量之间的相关关系还是可以接受的，可以进行下面的回归分析过程。

3．对数据进行回归分析的结果

图 10.15 是对数据进行回归分析的结果。

```
. regress TC Q PL PF PK

      Source |       SS       df       MS              Number of obs =     145
-------------+------------------------------           F(  4,   140) =  418.12
       Model |  52064.6433     4  13016.1608           Prob > F      =  0.0000
    Residual |  4358.19481   140   31.129963           R-squared     =  0.9228
-------------+------------------------------           Adj R-squared =  0.9206
       Total |  56422.8381   144  391.825265           Root MSE      =  5.5794

------------------------------------------------------------------------------
          TC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           Q |   .0063951   .0001629    39.26   0.000      .006073    .0067171
          PL |   5.655183    2.17636     2.60   0.010     1.352402    9.957964
          PF |     .20784   .0640999     3.24   0.001      .081111     .334569
          PK |   .0284415   .0265049     1.07   0.285    -.0239601    .0808431
       _cons |  -22.22098    6.58745    -3.37   0.001    -35.24472   -9.197235
------------------------------------------------------------------------------
```

图 10.15　回归分析的结果

从上述分析结果中，可以得到很多信息。可以看出共有 145 个样本参与了分析，模型的 F 值(4, 140) = 418.12，P 值（Prob > F）= 0.0000，说明模型整体上是非常显著的。模型的可决系数（R-squared） = 0.9228，模型修正的可决系数（Adj R-squared） = 0.9206，说明模型的解释能力还是差强人意的。

变量 Q 的系数标准误是 0.0001629，t 值为 39.26，P 值为 0.000，系数是非常显著的，95%的置信区间为[0.006073, 0.0067171]。变量 PL 的系数标准误是 2.17636，t 值为 2.60，P 值为 0.010，系数是非常显著的，95%的置信区间为[1.352402, 9.957964]。变量 PF 的系数标准误是 0.0640999，t 值为 3.24，P 值为 0.001，系数是非常显著的，95%的置信区间为[0.081111, 0.334569]。变量 PK 的系数标准误是 0.0265049，t 值为 1.07，P 值为 0.285，系数是非常不显著的，95%的置信区间为[-0.0239601, 0.0808431]。常数项的系数标准误是 6.58745，t 值为-3.37，P 值为 0.001，系数也是非常显著的，95%的置信区间为[-35.24472, -9.197235]。

模型的回归方程是：

TC=0.0063951*Q+ 5.655183*PL+0.20784*PF+0.0284415*PK -22.22098

从上面的分析可以看出美国电力企业的总成本（TC）受到产量（Q）、工资率（PL）、燃料价格（PF）及资本租赁价格（PK）的影响，美国电力行业存在规模经济。

4．对变量的方差-协方差矩阵

图 10.16 是对变量的方差-协方差矩阵。

```
. vce

Covariance matrix of coefficients of regress model

        e(V) |          Q          PL          PF          PK       _cons
-------------+------------------------------------------------------------
           Q |  2.654e-08
          PL |  -.0000764   4.7365431
          PF |  1.564e-06   -.0508677    .0041088
          PK | -2.741e-07   .01376813  -.00034147   .00070251
       _cons |  .00010096  -10.248761   .04900993  -.14021374   43.394499
```

图 10.16　变量的方差-协方差矩阵

从图 10.16 中可以看出，变量的方差与协方差都不是很大，有些甚至是微不足道的。

5．对变量系数的假设检验结果

图 10.17 是对变量系数的假设检验结果。

```
. test Q PL PF PK

 ( 1)  Q = 0
 ( 2)  PL = 0
 ( 3)  PF = 0
 ( 4)  PK = 0

       F(  4,   140) =  418.12
            Prob > F =    0.0000
```

图 10.17　对变量系数的假设检验结果

从图 10.17 中可以看出，模型非常显著，在 5%的显著性水平上通过了检验。

6．对因变量的拟合值的预测

图 10.18 是对因变量的拟合值的预测。

	TC	Q	PL	PF	PK	yhat
1	.082	2	2.09	17.9	183	-1.463724
2	.661	3	2.05	35.1	174	1.635338
3	.99	4	2.05	35.1	171	1.556408
4	.315	4	1.83	32.2	166	-.4326756
5	.197	5	2.12	28.6	233	2.371079
6	.098	9	2.12	28.6	195	1.315882
7	.949	11	1.98	35.5	206	2.283899
8	.675	13	2.05	35.1	150	1.016692
9	.525	13	2.19	29.1	155	.7035854
10	.501	22	1.72	15	188	-3.888769
11	1.194	25	2.09	17.9	170	-1.686378
12	.67	25	1.68	39.7	167	.4405845
13	.349	35	1.81	22.6	213	-1.006046
14	.423	39	2.3	23.6	164	.6047806
15	.501	43	1.75	42.8	170	1.681187
16	.55	63	1.76	10.3	161	-5.145133
17	.795	68	1.98	35.5	210	2.762183
18	.664	81	2.29	28.5	158	1.664588
19	.705	84	2.19	29.1	156	1.186076
20	.903	73	1.75	42.8	176	2.043688
21	1.504	99	2.2	36.2	170	3.212399
22	1.615	101	1.66	33.4	192	.2151507
23	1.127	119	1.92	22.5	164	-1.261209
24	.718	120	1.77	21.3	175	-2.039642
25	2.414	122	2.09	17.9	180	-.7816421
26	1.13	130	1.82	38.9	176	1.993493

图 10.18　对因变量的拟合值的预测

关于因变量预测拟合值的意义已在上节论述过，此处不再重复讲解。

7．回归分析得到的残差序列

图 10.19 是回归分析得到的残差序列。

关于残差序列的意义已在上节论述过，此处不再重复讲解。

读者应该注意到在上面的模型中，PK 的系数是不显著的，下面把该变量剔除掉重新进行回归分析。图 10.20 是对数据进行新回归分析的结果。

	TC	Q	PL	PF	PK	yhat	e
1	.082	2	2.09	17.9	183	-1.463724	1.545724
2	.661	3	2.05	35.1	174	1.635376	-.9743376
3	.99	4	2.05	35.1	171	1.556408	-.5664082
4	.315	4	1.83	32.2	166	-.4326756	.7476756
5	.197	5	2.12	28.6	233	2.371079	-2.174079
6	.098	9	2.12	28.6	195	1.315882	-1.217882
7	.949	11	1.98	35.5	206	2.283899	-1.334899
8	.675	13	2.05	35.1	150	1.016692	-.3416921
9	.525	13	2.19	29.1	155	.7035854	-.1785854
10	.501	22	1.72	15	188	-3.888769	4.38977
11	1.194	25	2.09	17.9	170	-1.686378	2.880378
12	.67	25	1.68	39.7	167	.4405845	.2294155
13	.349	35	1.81	22.6	213	-1.006046	1.355046
14	.423	39	2.3	23.6	164	.6047806	-.1817806
15	.501	43	1.75	42.8	170	1.681187	-1.180187
16	.55	63	1.76	10.3	161	-5.145133	5.695133
17	.795	68	1.98	35.5	210	2.762183	-1.967183
18	.664	81	2.29	28.5	158	1.664588	-1.000588
19	.705	84	2.19	29.1	156	1.186076	-.4810759
20	.903	73	1.75	42.8	176	2.043688	-1.140687
21	1.504	99	2.2	36.2	170	3.212398	-1.708398
22	1.615	101	1.66	33.4	192	.2151507	1.399849
23	1.127	119	1.92	22.5	164	-1.261208	2.388208
24	.718	120	1.77	21.3	175	-2.039642	2.757642
25	2.414	122	2.09	17.9	180	-.7816421	3.195642
26	1.13	130	1.82	38.9	176	1.993493	-.8634926
27	.992	138	1.8	20.2	202	-1.215579	2.207579

图 10.19　残差序列

```
. regress TC Q PL PF

      Source |       SS           df       MS      Number of obs   =       145
-------------+----------------------------------   F(3, 141)       =    556.52
       Model |  52028.7981         3  17342.9327   Prob > F        =    0.0000
    Residual |  4394.04007       141  31.1634048   R-squared       =    0.9221
-------------+----------------------------------   Adj R-squared   =    0.9205
       Total |  56422.8381       144  391.825265   Root MSE        =    5.5824

------------------------------------------------------------------------------
          TC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           Q |   .0064062   .0001627    39.38   0.000     .0060846    .0067277
          PL |   5.097772   2.114594     2.41   0.017     .9173653    9.278179
          PF |   .2216648   .0628256     3.53   0.001     .0974629    .3458667
       _cons |  -16.54434    3.92757    -4.21   0.000    -24.30888   -8.779805
------------------------------------------------------------------------------
```

图 10.20　新回归分析的结果

从上述分析结果中，可以看出模型整体依旧是非常显著的。模型的可决系数以及修正的可决系数（Adj R-squared）变化不大，说明模型的解释能力几乎没变。其他变量（包括常数项的系数）都非常显著，模型接近完美。可以把回归结果作为最终的回归模型方程，即：

TC=0.0064062*Q+ 5.097772*PL+0.2216648 *PF -16.54434

从上面的分析可以看出美国电力企业的总成本（TC）受到产量（Q）、工资率（PL）、燃料价格（PF）的影响，总成本随着这些变量的升高而升高、降低而降低。值得注意的是产量的增加引起总成本的相对变化是很小的，所以从经济意义上讲，美国的电力行业存在规模经济。

10.2.5　案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0

的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

1．延伸 1：在回归方程中不包含常数项

以本例为例进行说明，回归分析操作命令可以相应地修改为：

```
regress TC Q PL PF,nocon
```

在命令窗口输入命令并按回车键进行确认，结果如图 10.21 所示。

```
. regress TC Q PL PF,nocon

      Source |       SS       df       MS              Number of obs =     145
-------------+------------------------------           F(  3,   142) =  726.13
       Model |  75890.8019     3   25296.934           Prob > F      =  0.0000
    Residual |  4947.00303   142  34.8380495           R-squared     =  0.9388
-------------+------------------------------           Adj R-squared =  0.9375
       Total |   80837.805   145  557.502103           Root MSE      =  5.9024

------------------------------------------------------------------------------
          TC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           Q |   .0064558   .0001715    37.64   0.000     .0061167    .0067949
          PL |  -2.955539   .9553464    -3.09   0.002    -4.844079      -1.067
          PF |   .2011095   .0662258     3.04   0.003     .0701937    .3320253
------------------------------------------------------------------------------
```

图 10.21 延伸 1 分析结果图

从上述分析结果中，模型整体的显著程度依旧非常高。模型的可决系数（R-squared）及修正的可决系数略有上升，模型的解释能力更加强大。

模型的回归方程变为：

TC=0.0064558*Q-2.955539 *PL+0.2011095 *PF

值得注意的是，PL 的系数值竟然变为了负值，这说明 PL 的升高反而会带来总成本的降低，显然是不符合生活常识的，所以，该模型不可接受。

2．延伸 2：限定参与回归的样本范围

以本例为例进行说明，例如我们只对产量高于 100 的样本进行回归分析，操作命令可以相应地修改为：

```
regress TC Q PL PF if Q>=100
```

在命令窗口输入命令并按回车键进行确认，结果如图 10.22 所示。

```
. regress TC Q PL PF if Q>=100

      Source |       SS       df       MS              Number of obs =     124
-------------+------------------------------           F(  3,   120) =  450.85
       Model |  48385.1545     3  16128.3848           Prob > F      =  0.0000
    Residual |  4292.77683   120  35.7731402           R-squared     =  0.9185
-------------+------------------------------           Adj R-squared =  0.9165
       Total |  52677.9313   123  428.275864           Root MSE      =  5.9811

------------------------------------------------------------------------------
          TC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           Q |   .0064214    .000183    35.08   0.000      .006059    .0067839
          PL |    4.94541   2.457119     2.01   0.046     .0804852    9.810335
          PF |   .2674785   .0774243     3.45   0.001     .1141838    .4207732
       _cons |  -17.48977   4.418223    -3.96   0.000    -26.23755   -8.741999
------------------------------------------------------------------------------
```

图 10.22 延伸 2 分析结果图

关于结果的分析与前面类似，限于篇幅，这里不再赘述。

3．延伸 3：自动剔除不显著变量

在前面的分析过程中是采取逐步手动剔除不显著变量的方式得到了最终的回归模型，但是如果变量很多而且存在很多不显著的变量时，这个过程就显得非常复杂。那么有没有一种自动剔除不显著变量，直接得到最终模型方程的 Stata 操作方法呢？答案是肯定的。Stata 14.0 提供了 sw regress 命令来满足这一需要。这一命令的操作原理是不断迭代，最终使得所有变量系数的显著性达到设定的显著性水平。在首次迭代时，所有的变量都进入模型参与分析，然后每一步迭代都去掉 P 值最高或者说显著性最弱的变量。最终使得所有保留下来的变量的概率值都处于保留概率之下。以本例为例，如果设定显著性水平为 0.05，那么操作命令就应该是：

```
sw regress  TC Q PL PF PK,pr(0.05)
```

在命令窗口输入命令并按回车键进行确认，结果如图 10.23 所示。

```
. sw regress  TC Q PL PF PK,pr(0.05)
                      begin with full model
p = 0.2851 >= 0.0500  removing PK

      Source |       SS           df       MS      Number of obs   =       145
-------------+----------------------------------   F(  3,    141)   =    556.52
       Model |  52028.7981         3  17342.9327   Prob > F        =    0.0000
    Residual |  4394.04007       141  31.1634048   R-squared       =    0.9221
-------------+----------------------------------   Adj R-squared   =    0.9205
       Total |  56422.8381       144  391.825265   Root MSE        =    5.5824

------------------------------------------------------------------------------
          TC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           Q |   .0064062   .0001627    39.38   0.000     .0060846    .0067277
          PL |   5.097772   2.114594     2.41   0.017     .9173653    9.278179
          PF |   .2216648   .0628256     3.53   0.001     .0974629    .3458667
       _cons |  -16.54434    3.92757    -4.21   0.000    -24.30888   -8.779805
------------------------------------------------------------------------------
```

图 10.23　延伸 3 分析结果图

可以发现上述结果与前面逐步手动操作得到的结果一致。至于结果的详细解读，限于篇幅，这里不再赘述。

10.3　本章习题

（1）表 10.3 给出了 1955 年 145 家美国电力企业的总成本（TC）与产量（Q）的相关数据。试以总成本为因变量，以产量为自变量，利用简单回归分析方法研究其间的关系。

表 10.3　习题数据

编号	TC/百万美元	Q/千瓦时
1	0.082	2
2	0.661	3
3	0.990	4
4	0.315	4
5	0.197	5
6	0.098	9
…	…	…

（续表）

编号	TC/百万美元	Q/千瓦时
143	73.050	11796
144	139.422	14359
145	119.939	16719

（2）使用如表 10.4 所示的数据来估计教育投资的回报率。各变量说明如下：lw80（1980 年工人工资的对数值），s80（1980 年时工人的受教育年限），expr80（1980 年时工人的工龄），tenure80（1980 年时工人在现单位的工作年限），iq（智商），med（母亲的教育年限），kww（在 knowledge of the World of Work 测试中的成绩），mrt（婚姻虚拟变量，已婚=1），age（年龄）。模型说明：以 lw80 为因变量，以 s80、expr80、tenure80、iq 为自变量进行多重线性回归分析。

表 10.4　习题数据

mrt	med	iq	kww	age	s80	expr80	tenure80	lw80
0	8	93	35	19	12	10.64	2	6.64
0	14	119	41	23	18	11.37	16	6.69
0	14	108	46	20	14	11.03	9	6.72
0	12	96	32	18	12	13.09	7	6.48
1	6	74	27	26	11	14.40	5	6.33
0	8	91	24	16	10	13.43	0	6.40
…	…	…	…	…	…	…	…	…
1	12	101	38	25	12	10.59	5	6.47
1	7	100	33	23	12	9.00	3	6.17
1	8	102	32	19	13	9.83	3	7.09

第 11 章　Stata 回归诊断与应对

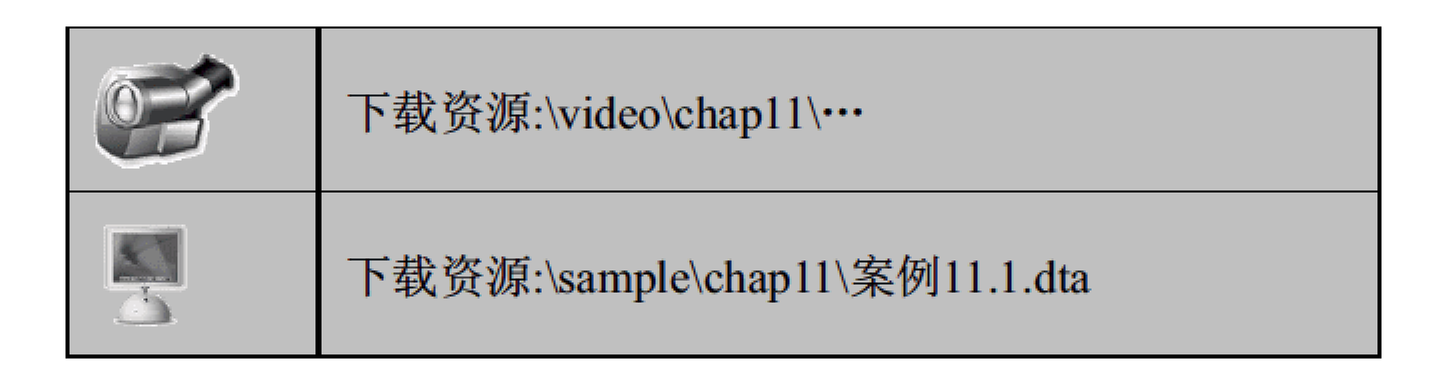
在上一章中，简要介绍了最小二乘线性回归，这种方法可以满足大部分的研究需要。但是这种分析方法的有效性建立在变量无异方差、无自相关、无多重共线性的基础之上。现实生活中很多数据是不满足这些条件的，那就需要用到将在本章中介绍的回归诊断与应对方法。本章的内容包括 3 部分，分别是异方差检验与应对、自相关检验与应对、多重共线性检验与应对等方法在实例中的具体应用。

11.1　实例一——异方差检验与应对

11.1.1　异方差检验与应对的功能与意义

在标准的线性回归模型中，有一个基本假设：整个总体同方差（也就是因变量的变异）不随自身预测值以及其他自变量的值的变化而变化。然而，在实际问题中这一假设条件往往不被满足，会出现异方差（Heteroskedasticity）的情况，如果继续采用标准的线性回归模型，就会使结果偏向于变异较大的数据，从而发生较大的偏差，所以在进行回归分析时往往需要检验变量的异方差，从而提出针对性的解决方案。常用的用于判断数据是否存在异方差的检验方法有绘制残差序列图、怀特检验、BP 检验等，解决异方差的方法有使用稳健的标准差进行回归以及使用加权最小二乘回归分析方法进行回归等。

11.1.2　相关数据来源



【例 11.1】某著名足球俱乐部拥有自己的一套球员评价体系，他们搜集并整理了其中 145 名球员的相关数据，如表 11.1 所示。表中的内容包括球员的身价、身体情况、精神情况、能力情况、潜力情况 5 部分的内容，试使用球员身价作为因变量，以球员的身体情况、精神情况、能力情况、潜力情况作为自变量，对这些数据使用最小二乘回归分析的方法进行研究，并进行异方差检验，最终建立合适的回归方程模型用于描述变量之间的关系。

表 11.1　某足球俱乐部搜集整理的 145 名球员的相关数据

编号	球员身价	身体情况	精神情况	能力情况	潜力情况
1	4.406 719	0.693 147	5.342 334	5.187 386	5.209 486
2	6.493 754	1.098 612	5.323 01	5.860 786	5.159 055
3	6.897 705	1.386 294	5.323 01	5.860 786	5.141 664
4	5.752 573	1.386 294	5.209 486	5.774 552	5.111 988
5	5.283 204	1.609 438	5.356 586	5.655 992	5.451 039
6	4.584 968	2.197 225	5.356 586	5.655 992	5.273
…	…	…	…	…	…
142	11.114 24	9.348 1	5.411 646	5.579 73	5.017 28
143	11.198 9	9.375 516	5.356 586	5.655 992	4.997 212
144	11.845 26	9.572 132	5.442 418	5.814 131	5.356 586
145	11.694 74	9.724 301	5.438 079	5.463 832	5.087 596

11.1.3　Stata 分析过程

在利用 Stata 进行分析之前，要把数据录入到 Stata 中。本例中有 5 个变量，分别为球员的身价、身体情况、精神情况、能力情况、潜力情况。我们把这 5 个变量分别设定为 V1~V5，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作在第 1 章中已有详细讲述。录入完成后数据如图 11.1 所示。

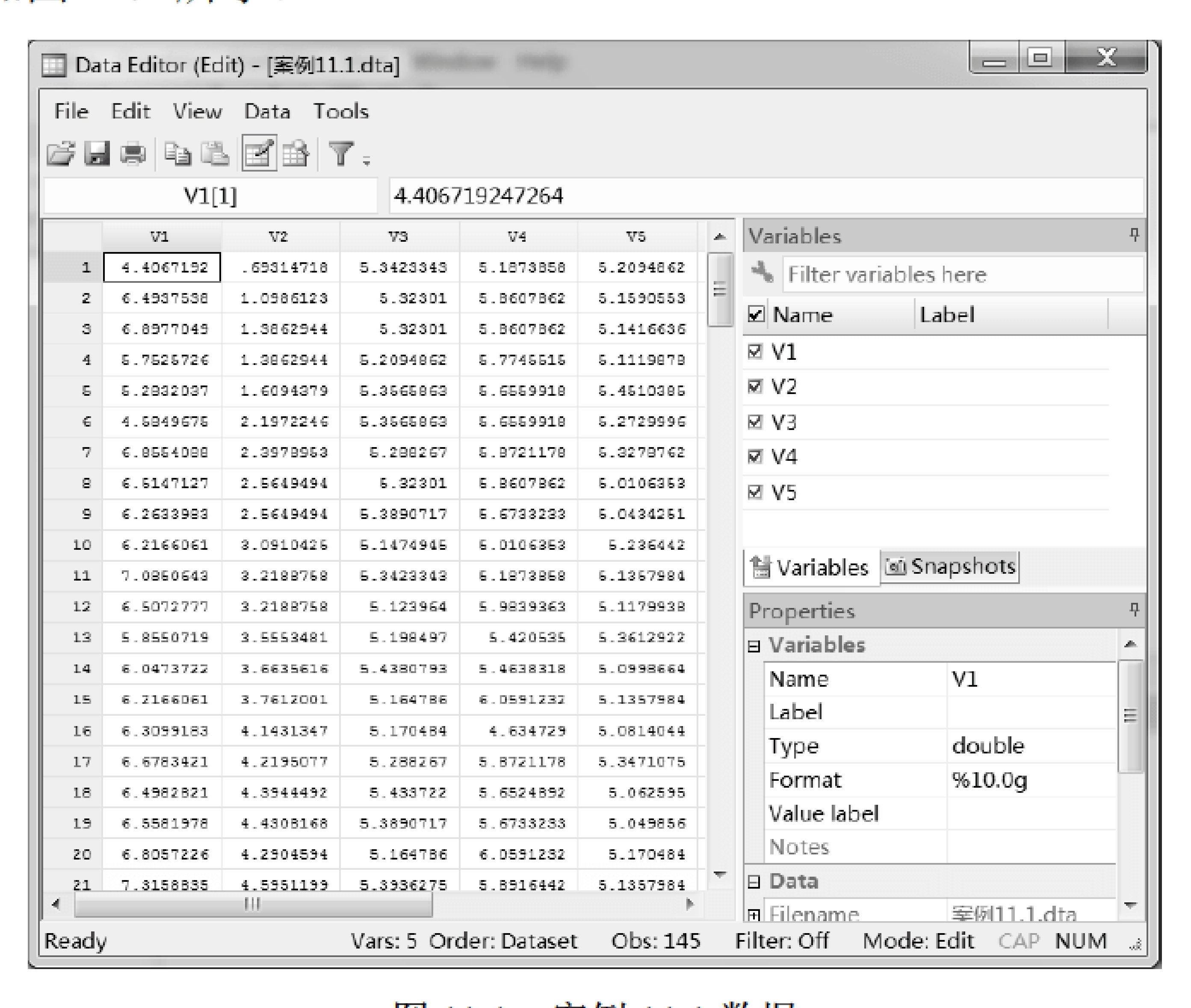


图 11.1　案例 11.1 数据

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入如下命令。

- summarize　V1 V2 V3 V4 V5,detail：本命令旨在对数据进行描述性分析，从总体上探

索数据特征，观测其是否存在极端数据或者变量间的量纲差距过大，从而可能会对回归分析结果造成不利影响。

- correlate V1 V2 V3 V4 V5：本命令旨在对数据进行相关性分析，旨在探索变量之间尤其是因变量与各个自变量之间的相关性关系，该步骤是进行回归分析前的必要准备。
- regress V1 V2 V3 V4 V5：本命令旨在对数据进行回归分析，探索自变量对因变量的影响情况。
- vce：本命令旨在获得变量的方差-协方差矩阵。
- test V2 V3 V4 V5：本命令旨在检验回归分析获得的各个自变量系数的显著性。
- predict yhat：本命令旨在获得因变量的拟合值。
- predict e,resid：本命令旨在获得回归模型的估计残差。
- rvfplot：本命令旨在绘制残差与回归得到的拟合值的散点图，从而探索数据是否存在异方差。
- rvpplot V2：本命令旨在绘制残差与解释变量 V2 的散点图，从而探索数据是否存在异方差。
- estat imtest,white：本命令为怀特检验，旨在检验数据是否存在异方差。
- estat hettest,iid：本命令为 BP 检验，旨在使用得到的拟合值来检验数据是否存在异方差。
- estat hettest,rhs iid：本命令为 BP 检验，旨在使用方程右边的解释数据来检验变量是否存在异方差。
- estat hettest V2,rhs iid：本命令为 BP 检验，旨在使用指定的解释数据 V2 来检验变量是否存在异方差。
- regress V1 V2 V3 V4 V5,robust：本命令为采用稳健的标准差对数据进行回归分析，克服数据的异方差对最小二乘回归分析造成的不利影响。

03 设置完毕后，按键盘上的回车键，等待输出结果。

11.1.4 结果分析

在 Stata 14.0 主界面的结果窗口可以看到如图 11.2~图 11.15 所示的分析结果。

1. 对数据进行描述性分析的结果

图 11.2 是对数据进行描述性分析的结果。关于这一分析过程对于回归分析的重要意义已在前面章节中论述过，此处不再重复讲解。

```
. summarize  V1 V2 V3 V4 V5,detail

                             V1
-------------------------------------------------------------
      Percentiles      Smallest
 1%     4.584967       4.406719
 5%     6.216606       4.584967
10%     6.558198       5.283204       Obs                 145
25%     7.775696       5.752573       Sum of Wgt.         145

50%      8.81789                      Mean           8.632419
                        Largest       Std. Dev.      1.421723
75%     9.556197       11.15451
90%     10.38338        11.1989       Variance       2.021297
95%     10.71206       11.69474       Skewness      -.4086256
99%     11.69474       11.84526       Kurtosis       3.064497

                             V2
-------------------------------------------------------------
      Percentiles      Smallest
 1%     1.098612       .6931472
 5%     2.564949       1.098612
10%       3.7612       1.386294       Obs                 145
25%     5.631212       1.386294       Sum of Wgt.         145

50%     7.011214                      Mean           6.556651
                        Largest       Std. Dev.      1.912792
75%     7.826842         9.3481
90%     8.668884       9.375516       Variance       3.658775
95%     9.064389       9.572132       Skewness      -.9612785
99%     9.572132       9.724301       Kurtosis        3.65205

                             V3
-------------------------------------------------------------
      Percentiles      Smallest
 1%     4.976734       4.976734
 5%     5.043425       4.976734
10%     5.123964       5.023881       Obs                 145
25%     5.170484       5.023881       Sum of Wgt.         145

50%      5.31812                      Mean           5.276838
                        Largest       Std. Dev.      .1233593
75%     5.389072       5.446737
90%     5.438079       5.446737       Variance       .0152175
95%     5.442418       5.446737       Skewness       -.429873
99%     5.446737       5.446737       Kurtosis       2.179193

                             V4
-------------------------------------------------------------
      Percentiles      Smallest
 1%     4.634729       4.634729
 5%     4.634729       4.634729
10%     4.859812       4.634729       Obs                 145
25%     5.361292       4.634729       Sum of Wgt.         145

50%     5.594711                      Mean           5.511444
                        Largest       Std. Dev.      .3589003
75%     5.774552       5.983936
90%     5.860786       6.059123       Variance       .1288094
95%     5.891644       6.059123       Skewness      -1.126801
99%     6.059123       6.059123       Kurtosis       3.747527

                             V5
-------------------------------------------------------------
      Percentiles      Smallest
 1%     4.962845       4.927254
 5%     5.043425       4.962845
10%     5.056246       4.969813       Obs                 145
25%     5.087596       4.997212       Sum of Wgt.         145

50%     5.135798                      Mean           5.156777
                        Largest       Std. Dev.      .1003897
75%     5.209486         5.4161
90%     5.308268         5.4161       Variance       .0100781
95%     5.356586        5.42495       Skewness       .7363024
99%      5.42495       5.451038       Kurtosis       3.296593
```

图 11.2　案例 11.1 描述性分析的结果

在如图 11.2 所示的分析结果中，可以得到很多信息，包括百分位数、4 个最小值、4 个最大值、平均值、标准差、偏度、峰度等。

（1）百分位数（Percentiles）

可以看出变量 V1 的第 1 个四分位数(25%)是 7.775696，第 2 个四分位数(50%)是 8.81789，第 3 个四分位数（75%）是 9.556197；变量 V2 的第 1 个四分位数（25%）是 5.631212，第 2 个四分位数（50%）是 7.011214，第 3 个四分位数（75%）是 7.826842；变量 V3 的第 1 个四分位数（25%）是 5.170484，第 2 个四分位数（50%）是 5.31812，第 3 个四分位数（75%）是 5.389072；变量 V4 的第 1 个四分位数(25%)是 5.361292，第 2 个四分位数(50%)是 5.594711，第 3 个四分位数（75%）是 5.774552；变量 V5 的第 1 个四分位数（25%）是 5.087596，第 2 个四分位数（50%）是 5.135798，第 3 个四分位数（75%）是 5.209486。

（2）4 个最小值（Smallest）

变量 V1 最小的 4 个数据值分别是 4.406719、4.584967、5.283204、5.752573。

变量 V2 最小的 4 个数据值分别是 0.6931472、1.098612、1.386294、1.386294。

变量 V3 最小的 4 个数据值分别是 4.976734、4.976734、5.023881、5.023881。

变量 V4 最小的 4 个数据值分别是 4.634729、4.634729、4.634729、4.634729。

变量 V5 最小的 4 个数据值分别是 4.927254、4.962845、4.969813、4.997212。

（3）4 个最大值（Largest）

变量 V1 最大的 4 个数据值分别是 11.15451、11.1989、11.69474、11.84526。

变量 V2 最大的 4 个数据值分别是 9.3481、9.375516、9.572132、9.724301。

变量 V3 最大的 4 个数据值分别是 5.446737、5.446737、5.446737、5.446737。

变量 V4 最大的 4 个数据值分别是 5.983936、6.059123、6.059123、6.059123。

变量 V5 最大的 4 个数据值分别是 5.4161、5.4161、5.42495、5.451038。

（4）平均值（Mean）和标准差（Std. Dev）

变量 V1 的平均值为 8.632419，标准差是 1.421723。

变量 V2 的平均值为 6.556651，标准差是 1.912792。

变量 V3 的平均值为 5.276838，标准差是 0.1233593。

变量 V4 的平均值为 5.511444，标准差是 0.3589003。

变量 V5 的平均值为 5.156777，标准差是 0.1003897。

（5）偏度（Skewness）和峰度（Kurtosis）

变量 V1 的偏度为-0.4086256，为负偏度但不大。

变量 V2 的偏度为-0.9612785，为负偏度但不大。

变量 V3 的偏度为-0.429873，为负偏度但不大。

变量 V4 的偏度为-1.126801，为负偏度但不大。

变量 V5 的偏度为 0.7363024，为正偏度但不大。

变量 V1 的峰度为 3.064497，有一个比正态分布略长的尾巴。

变量 V2 的峰度为 3.65205，有一个比正态分布略长的尾巴。

变量 V3 的峰度为 2.179193，有一个比正态分布略短的尾巴。

变量 V4 的峰度为 3.747527，有一个比正态分布略长的尾巴。

变量 V5 的峰度为 3.296593，有一个比正态分布略长的尾巴。

综上所述，数据的总体质量还是可以的，没有极端异常值，变量间的量纲差距、变量的偏度、峰度也是可以接受的，可以进行下一步的分析。

2. 对数据进行相关性分析的结果

图 11.3 是对数据进行相关性分析的结果。关于这一分析过程对于回归分析的重要意义已在前面章节中论述过，此处不再重复讲解。

```
. correlate  V1 V2 V3 V4 V5
(obs=145)

             |       V1       V2       V3       V4       V5
-------------+---------------------------------------------
          V1 |   1.0000
          V2 |   0.9542   1.0000
          V3 |   0.1174   0.0422   1.0000
          V4 |  -0.0455  -0.1609   0.3319   1.0000
          V5 |  -0.1042  -0.0988  -0.1865   0.1309   1.0000
```

图 11.3　案例 11.1 相关性分析的结果

在图 11.3 中，V1 与各个自变量之间的相关关系还是可以接受的，可以进入下面的回归分析过程。

3．对数据进行回归分析的结果

图 11.4 是对数据进行回归分析的结果。

```
. regress  V1 V2 V3 V4 V5

      Source |       SS       df       MS              Number of obs =     145
-------------+------------------------------           F(  4,   140) =  437.69
       Model |  269.514818     4  67.3787045           Prob > F      =  0.0000
    Residual |  21.5520082   140  .153942915           R-squared     =  0.9260
-------------+------------------------------           Adj R-squared =  0.9238
       Total |  291.066826   144  2.0212974            Root MSE      =  .39236

------------------------------------------------------------------------------
          V1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          V2 |   .7203941   .0174664    41.24   0.000      .685862    .7549262
          V3 |   .4363412   .2910476     1.50   0.136    -.1390756    1.011758
          V4 |    .426517   .1003691     4.25   0.000     .2280818    .6249521
          V5 |  -.2198884   .3394286    -0.65   0.518     -.890957    .4511803
       _cons |   .3897354   2.455817     0.16   0.874    -4.465547    5.245018
------------------------------------------------------------------------------
```

图 11.4　案例 11.1 回归分析的结果

从上述分析结果中，可以看出共有 145 个样本参与了分析，模型的 F 值(4, 140) = 437.69，P 值（Prob > F） = 0.0000，说明模型整体上是非常显著的。模型的可决系数（R-squared）= 0.9260，模型修正的可决系数（Adj R-squared） = 0.9238，说明模型的解释能力非常不错。

模型的回归方程是：

V1=0.7203941*V2+0.4363412*V3+0.426517*V4-0.2198884*V5+0.3897354

变量 V2 的系数标准误是 0.0174664，t 值为 41.24，P 值为 0.000，系数是非常显著的，95%的置信区间为[0.685862, 0.7549262]。变量 V3 的系数标准误是 0.2910476，t 值为 1.5，P 值为 0.136，系数的显著程度不高，95%的置信区间为[-0.1390756, 1.011758]。变量 V4 的系数标准误是 0.1003691，t 值为 4.25，P 值为 0.000，系数是非常显著的，95%的置信区间为[0.2280818, 0.6249521]。变量 V5 的系数标准误是 0.3394286，t 值为-0.65，P 值为 0.518，系数是非常不显著的，95%的置信区间为[-0.890957, 0.4511803]。常数项的系数标准误是 2.455817，t 值为 0.16，P 值为 0.874，系数也是非常不显著的，95%的置信区间为[-4.465547, 5.245018]。

从上面的分析可以看出，球员的身价与其身体情况、精神情况、能力情况之间是一种正向联动的变化关系，这在事实上也是可以接受的，但是球员的潜力情况对身价影响的显著性很低，而且是一种负值关系，这可能是因为球员的潜力情况本身就很难衡量，或其预测存在很大偏差所致。

4．对变量的方差-协方差矩阵

图 11.5 是对变量的方差-协方差矩阵。

```
. vce

Covariance matrix of coefficients of regress model

        e(V) |         V2          V3          V4          V5       _cons
-------------+------------------------------------------------------------
          V2 |  .00030508
          V3 | -.00045476   .08470871
          V4 |  .00031477  -.01094123   .01007397
          V5 |  .00032263   .02367632  -.00662951   .11521179
       _cons | -.00299912  -.50580392   .03433611  -.68463476   6.0310381
```

图 11.5　变量的方差-协方差矩阵

从图 11.5 中可以看出，各个自变量的方差与协方差都不是很大。

5．对变量系数的假设检验结果

图 11.6 是对变量系数的假设检验结果。

```
. test V2 V3 V4 V5

 ( 1)  V2 = 0
 ( 2)  V3 = 0
 ( 3)  V4 = 0
 ( 4)  V5 = 0

       F(  4,   140) =  437.69
            Prob > F =    0.0000
```

图 11.6　对变量系数的假设检验结果

从图 11.6 中可以看出，模型非常显著，在 5%的显著性水平上通过了检验。

6．对因变量的拟合值的预测

图 11.7 是对因变量的拟合值的预测。

	V1	V2	V3	V4	V5	yhat
1	4.4067192	.69314718	5.3423343	5.1873858	5.2094862	4.287158
2	6.4937538	1.0986123	5.32301	5.8607862	5.1590553	4.869126
3	6.8977049	1.3862944	5.32301	5.8607862	5.1416636	5.080195
4	5.7525726	1.3862944	5.2094862	5.7745515	5.1119878	5.000405
5	5.2832037	1.6094379	5.3565863	5.6559918	5.4510385	5.100221
6	4.5849675	2.1972246	5.3565863	5.6559918	5.2729996	5.562808
7	6.8554088	2.3978953	5.288267	5.8721178	5.3278762	5.757674
8	6.5147127	2.5649494	5.32301	5.8607862	5.0106353	5.958103
9	6.2633983	2.5649494	5.3890717	5.6733233	5.0434251	5.899762
10	6.2166061	3.0910425	5.1474945	5.0106353	5.236442	5.848256
11	7.0850643	3.2188758	5.3423343	5.1873858	5.1357984	6.122881
12	6.5072777	3.2188758	5.123964	5.9839363	5.1179938	6.371254
13	5.8550719	3.5553481	5.198497	5.420535	5.3612922	6.35237
14	6.0473722	3.6635616	5.4380793	5.4638318	5.0998664	6.610817
15	6.2166061	3.7612001	5.164786	6.0591232	5.1357984	6.807907
16	6.3099183	4.1431347	5.170484	4.634729	5.0814044	6.489969
17	6.6783421	4.2195077	5.288267	5.8721178	5.3471075	7.065723
18	6.4982821	4.3944492	5.433722	5.6524892	5.062595	7.224104
19	6.5581978	4.4308168	5.3890717	5.6733233	5.049856	7.242507
20	6.8057226	4.2904594	5.164786	6.0591232	5.170484	7.181555
21	7.3158835	4.5951199	5.3936275	5.8916442	5.1357984	7.437078
22	7.3870902	4.6151205	5.1119878	5.811141	5.2574954	7.2675
23	7.0273145	4.7791235	5.2574954	5.4161004	5.0998664	7.315307
24	6.5764696	4.7874917	5.1761497	5.3612922	5.164786	7.248189
25	7.7890404	4.804021	5.3423343	5.1873858	5.1929569	7.252242
26	7.0299729	4.8675345	5.2040067	5.9635793	5.170484	7.573639
27	6.8997231	4.9272537	5.1929569	5.3082677	5.3082677	7.302041
28	7.3485875	5.0039463	5.2574954	5.4161004	5.42495	7.405786

图 11.7　对因变量的拟合值的预测

因变量预测拟合值是根据自变量的值和得到的回归方程计算出来的，主要用于预测未来。在图 11.7 中，可以看到 yhat 的值与 V1 的值是比较相近的，所以拟合的回归模型还是不错的。

7．回归分析得到的残差序列

图 11.8 是回归分析得到的残差序列。

	V1	V2	V3	V4	V5	yhat	e
1	4.4067192	.69314718	5.3423343	5.1873858	5.2094862	4.287158	.1195615
2	6.4937538	1.0986123	5.32301	5.8607862	5.1590553	4.869126	1.624628
3	6.8977049	1.3862944	5.32301	5.8607862	5.1416636	5.080195	1.81751
4	5.7525726	1.3862944	5.2094862	5.7745515	5.1119878	5.000405	.752168
5	5.2832037	1.6094379	5.3565863	5.6559918	5.4510385	5.100221	.182983
6	4.5849675	2.1972246	5.3565863	5.6559918	5.2729996	5.562808	-.9778399
7	6.8554088	2.3978953	5.288267	5.8721178	5.3278762	5.757674	1.097735
8	6.5147127	2.5649494	5.32301	5.8607862	5.0106353	5.958103	.5566101
9	6.2633983	2.5649494	5.3890717	5.6733233	5.0434251	5.899762	.3636364
10	6.2166061	3.0910425	5.1474945	5.0106353	5.236442	5.848256	.3683498
11	7.0850643	3.2188758	5.3423343	5.1873858	5.1357984	6.122881	.9621836
12	6.5072777	3.2188758	5.123964	5.9839363	5.1179938	6.371254	.1360236
13	5.8550719	3.5553481	5.198497	5.420535	5.3612922	6.35237	-.4972979
14	6.0473722	3.6635616	5.4380793	5.4638318	5.0998664	6.610817	-.563445
15	6.2166061	3.7612001	5.164786	6.0591232	5.1357984	6.807907	-.591301
16	6.3099183	4.1431347	5.170484	4.634729	5.0814044	6.489969	-.1800509
17	6.6783421	4.2195077	5.288267	5.8721178	5.3471075	7.065723	-.3873815
18	6.4982821	4.3944492	5.433722	5.6524892	5.062595	7.224104	-.7258219
19	6.5581978	4.4308168	5.3890717	5.6733233	5.049856	7.242507	-.6843098
20	6.8057226	4.2904594	5.164786	6.0591232	5.170484	7.181555	-.3758329
21	7.3158835	4.5951199	5.3936275	5.8916442	5.1357984	7.437078	-.1211948
22	7.3870902	4.6151205	5.1119878	5.811141	5.2574954	7.2675	.1195903
23	7.0273145	4.7791235	5.2574954	5.4161004	5.0998664	7.315307	-.2879924
24	6.5764696	4.7874917	5.1761497	5.3612922	5.164786	7.248189	-.6717196
25	7.7890404	4.804021	5.3423343	5.1873858	5.1929569	7.252242	.5367989
26	7.0299729	4.8675345	5.2040067	5.9635793	5.170484	7.573639	-.5436665
27	6.8997231	4.9272537	5.1929569	5.3082677	5.3082677	7.302041	-.4023176
28	7.3485875	5.0039463	5.2574954	5.4161004	5.42495	7.405786	-.0571983

图 11.8　残差序列

关于残差序列的意义已在上节中论述过，此处不再重复讲解。

8．绘制散点图

图 11.9 是利用上面两步得到的残差与得到的拟合值绘制的散点图。

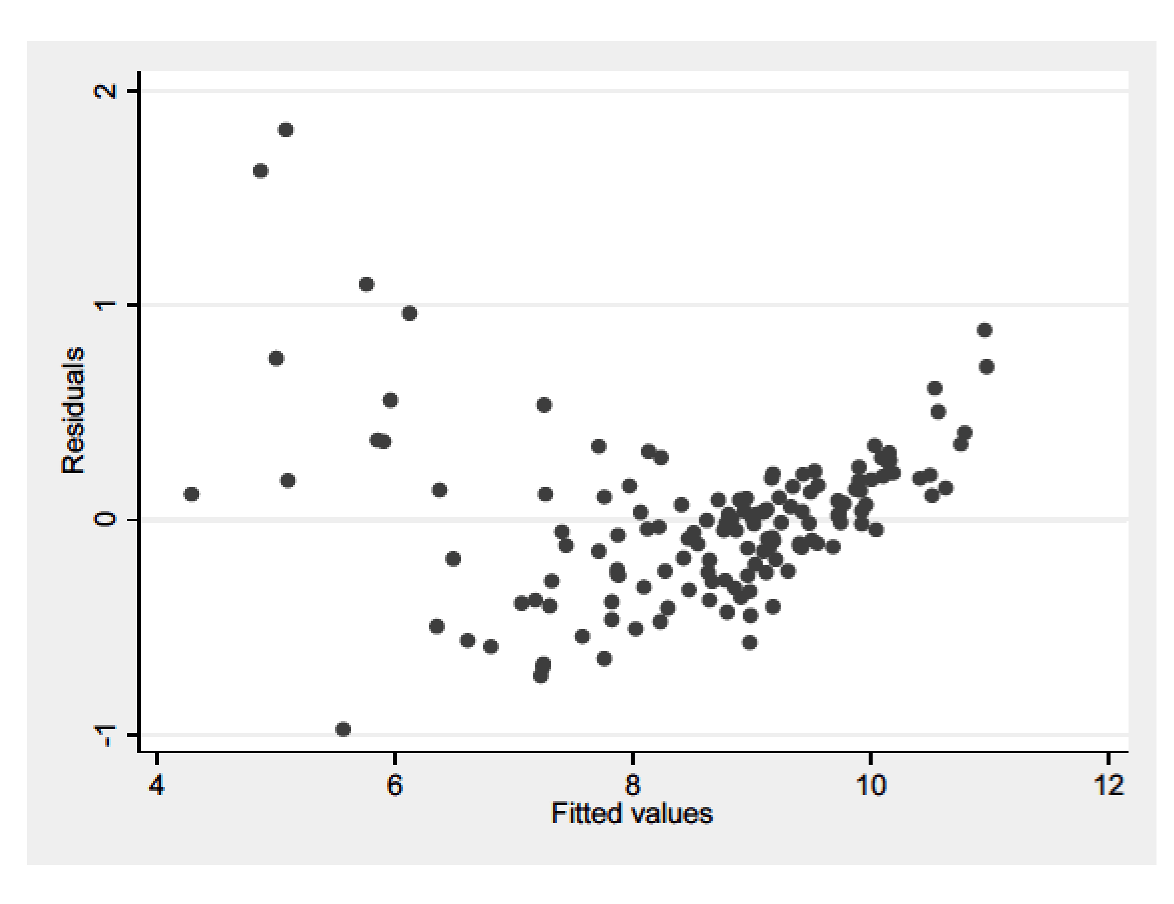


图 11.9　残差与拟合值的散点图

从图 11.9 中可以看出，残差随着拟合值的不同而有所不同，尤其是在拟合值较小（4~8）的时候，残差波动比较剧烈（并不是在 0 附近），所以，数据是存在异方差的。

图 11.10 是利用残差与自变量 V2 绘制的散点图。

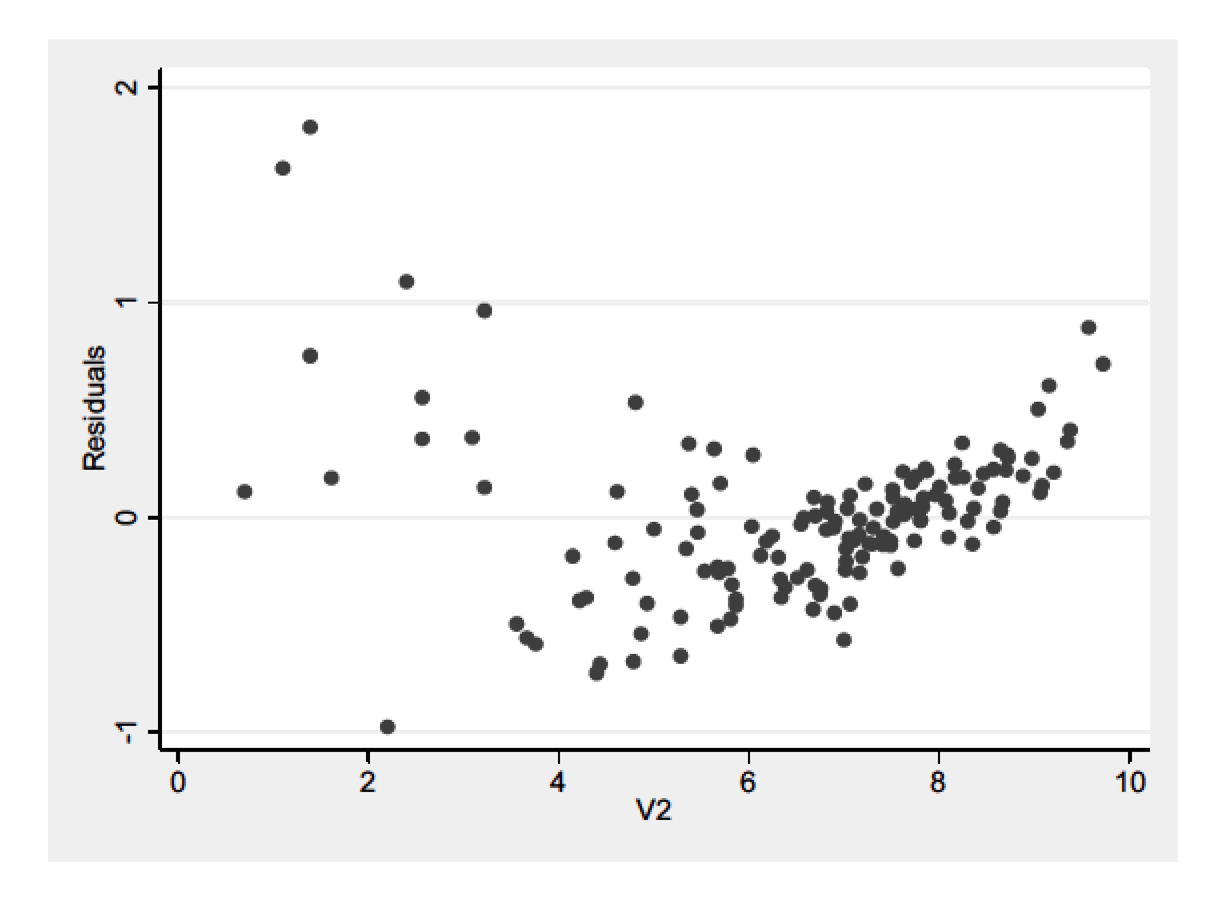


图 11.10　残差与自变量 V2 的散点图

从图 11.10 中可以看出，残差随着自变量 V2 值的不同而有所不同，尤其是在 V2 值较小（0~4）的时候，残差波动比较剧烈（并不是在 0 附近），所以，数据是存在异方差的。

9．怀特检验的检验结果

图 11.11 是怀特检验的检验结果。

```
. estat imtest,white

White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(14)      =     73.48
         Prob > chi2   =    0.0000

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df      p
---------------------+-----------------------------
  Heteroskedasticity |      73.48     14    0.0000
            Skewness |      22.34      4    0.0002
            Kurtosis |       2.62      1    0.1052
---------------------+-----------------------------
               Total |      98.45     19    0.0000
---------------------------------------------------
```

图 11.11　怀特检验的检验结果

怀特检验的原假设数据为同方差。从图 11.11 中可以看出，P 值为 0.0000，非常显著地拒绝了同方差的原假设，认为存在异方差。

10．BP 检验的检验结果

图 11.12~图 11.14 是 BP 检验的检验结果。其中，图 11.12 是使用得到的拟合值对数据进行异方差检验的结果，图 11.13 是使用方程右边的解释变量对数据进行异方差检验的结果，图 11.14 是使用指定的解释变量 V2 对数据进行异方差检验的结果。

```
. estat hettest,iid

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of V1

         chi2(1)      =    29.04
         Prob > chi2  =   0.0000

.
```

图 11.12　BP 检验的检验结果 1

```
. estat hettest,rhs iid

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: V2 V3 V4 V5

         chi2(4)      =    35.55
         Prob > chi2  =   0.0000
```

图 11.13　BP 检验的检验结果 2

```
. estat hettest V2,rhs iid

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: V2 V3 V4 V5

         chi2(4)      =    35.55
         Prob > chi2  =   0.0000
```

图 11.14　BP 检验的检验结果 3

BP 检验的原假设数据为同方差。从图 11.12~图 11.14 中可以看出，P 值均为 0.0000，非常显著地拒绝了同方差的原假设，认为存在异方差。

11. 回归分析的结果

图 11.15 是使用稳健的标准差对数据进行回归分析的结果。

```
. regress  V1 V2 V3 V4 V5,robust

Linear regression                                      Number of obs =      145
                                                       F(  4,   140) =   175.79
                                                       Prob > F      =   0.0000
                                                       R-squared     =   0.9260
                                                       Root MSE      =   .39236

------------------------------------------------------------------------------
             |               Robust
          V1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          V2 |   .7203941   .0325975    22.10   0.000      .655947    .7848411
          V3 |   .4363412   .2456358     1.78   0.078     -.049294    .9219764
          V4 |    .426517   .0754827     5.65   0.000     .2772836    .5757503
          V5 |  -.2198884   .3238121    -0.68   0.498    -.8600823    .4203056
       _cons |   .3897354    2.30735     0.17   0.866    -4.172019     4.95149
------------------------------------------------------------------------------
```

图 11.15　使用稳健的标准差对数据进行回归分析的结果

从上述分析结果中可以得到很多信息。可以看出模型的 F 值(4, 140) = 175.79，P 值（Prob > F）= 0.0000，说明模型整体上依旧是非常显著的。模型的可决系数（R-squared）为 0.9260，模型的解释能力依旧很高。

模型的回归方程没有发生变化，依旧是：

V1=0.7203941*V2+0.4363412*V3+0.426517*V4-0.2198884*V5+0.3897354

但是 V3、V5 等变量系数的显著性得到了一定程度的提高，这说明通过使用稳健的标准差进行回归分析，使回归模型得到了一定程度的改善。

11.1.5 案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

下面使用加权最小二乘回归分析方法解决数据的异方差问题。

以本例为例进行说明，操作命令如下。

- reg V1-V5：本命令旨在以 V1 为因变量，以 V2、V3、V4、V5 为自变量，对数据进行最小二乘回归分析。
- predict e,resid：本命令旨在估计上步回归分析得到的残差。
- gen ee=e^2：本命令旨在对残差数据进行平方变换，产生的新变量 ee 为残差的平方。
- gen lnee=log(ee)：本命令旨在对数据进行对数变换，产生的新变量 lnee 为上步得到残差平方的对数值。
- reg lnee V2,nocon：本命令旨在进行以上步得到的残差平方对数值为因变量，以 V2 为自变量，并且不包含常数项的最小二乘回归分析。
- predict yhat：本命令旨在预测上步进行的最小二乘回归产生的因变量的拟合值。
- gen yhathat=exp(yhat)：本命令旨在对因变量的拟合值进行指数变换，产生的新变量 yhathat 为 yhat 的指数值。
- reg V1 V2 V3 V4 V5 [aw=1/yhathat]：本命令旨在对数据进行以 V1 为因变量，以 V2、V3、V4、V5 为自变量，以 yhathat 的倒数为权重变量的加权最小二乘回归分析。

在命令窗口输入命令并按回车键进行确认，结果如图 11.16~图 11.23 所示。

图 11.16 是对数据进行回归分析的结果。

```
. reg V1-V5

      Source |       SS           df       MS      Number of obs   =       145
-------------+----------------------------------   F(  4,    140)  =    437.69
       Model |  269.514818         4  67.3787045   Prob > F        =    0.0000
    Residual |  21.5520082       140  .153942915   R-squared       =    0.9260
-------------+----------------------------------   Adj R-squared   =    0.9238
       Total |  291.066826       144   2.0212974   Root MSE        =    .39236

------------------------------------------------------------------------------
          V1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          V2 |   .7203941   .0174664    41.24   0.000      .685862    .7549262
          V3 |   .4363412   .2910476     1.50   0.136    -.1390756    1.011758
          V4 |    .426517   .1003691     4.25   0.000     .2280818    .6249521
          V5 |  -.2198884   .3394286    -0.65   0.518     -.890957    .4511803
       _cons |   .3897354   2.455817     0.16   0.874    -4.465547    5.245018
------------------------------------------------------------------------------
```

图 11.16 对数据进行回归分析的结果

对本结果的解读已在前面有所表述，此处限于篇幅不再赘述。

图 11.17 是回归分析得到的残差序列。

	V1	V2	V3	V4	V5	e
1	4.4067192	.69314718	5.3423343	5.1873858	5.2094862	.1195615
2	6.4937538	1.0986123	5.32301	5.8607862	5.1590553	1.624628
3	6.8977049	1.3862944	5.32301	5.8607862	5.1416636	1.81751
4	5.7525726	1.3862944	5.2094862	5.7745515	5.1119878	.752168
5	5.2832037	1.6094379	5.3565863	5.6559918	5.4510385	.182983
6	4.5849675	2.1972246	5.3565863	5.6559918	5.2729996	-.9778399
7	6.8554088	2.3978953	5.288267	5.8721178	5.3278762	1.097735
8	6.5147127	2.5649494	5.32301	5.8607862	5.0106353	.5566101
9	6.2633983	2.5649494	5.3890717	5.6733233	5.0434251	.3636364
10	6.2166061	3.0910425	5.1474945	5.0106353	5.236442	.3683498
11	7.0850643	3.2188758	5.3423343	5.1873858	5.1357984	.9621836
12	6.5072777	3.2188758	5.123964	5.9839363	5.1179938	.1360236
13	5.8550719	3.5553481	5.198497	5.420535	5.3612922	-.4972979
14	6.0473722	3.6635616	5.4380793	5.4638318	5.0998664	-.563445
15	6.2166061	3.7612001	5.164786	6.0591232	5.1357984	-.591301
16	6.3099183	4.1431347	5.170484	4.634729	5.0814044	-.1800509
17	6.6783421	4.2195077	5.288267	5.8721178	5.3471075	-.3873815
18	6.4982821	4.3944492	5.433722	5.6524892	5.062595	-.7258219
19	6.5581978	4.4308168	5.3890717	5.6733233	5.049856	-.6843098
20	6.8057226	4.2904594	5.164786	6.0591232	5.170484	-.3758329
21	7.3158835	4.5951199	5.3936275	5.8916442	5.1357984	-.1211948
22	7.3870902	4.6151205	5.1119878	5.811141	5.2574954	.1195903

图 11.17 回归分析得到的残差序列

图 11.18 是对残差序列进行平方变换后的结果。

	V1	V2	V3	V4	V5	e	ee
1	4.4067192	.69314718	5.3423343	5.1873858	5.2094862	.1195615	.014295
2	6.4937538	1.0986123	5.32301	5.8607862	5.1590553	1.624628	2.639415
3	6.8977049	1.3862944	5.32301	5.8607862	5.1416636	1.81751	3.303343
4	5.7525726	1.3862944	5.2094862	5.7745515	5.1119878	.752168	.5657567
5	5.2832037	1.6094379	5.3565863	5.6559918	5.4510385	.182983	.0334828
6	4.5849675	2.1972246	5.3565863	5.6559918	5.2729996	-.9778399	.956171
7	6.8554088	2.3978953	5.288267	5.8721178	5.3278762	1.097735	1.205022
8	6.5147127	2.5649494	5.32301	5.8607862	5.0106353	.5566101	.3098148
9	6.2633983	2.5649494	5.3890717	5.6733233	5.0434251	.3636364	.1322315
10	6.2166061	3.0910425	5.1474945	5.0106353	5.236442	.3683498	.1356816
11	7.0850643	3.2188758	5.3423343	5.1873858	5.1357984	.9621836	.9257973
12	6.5072777	3.2188758	5.123964	5.9839363	5.1179938	.1360236	.0185024
13	5.8550719	3.5553481	5.198497	5.420535	5.3612922	-.4972979	.2473052
14	6.0473722	3.6635616	5.4380793	5.4638318	5.0998664	-.563445	.3174703
15	6.2166061	3.7612001	5.164786	6.0591232	5.1357984	-.591301	.3496368
16	6.3099183	4.1431347	5.170484	4.634729	5.0814044	-.1800509	.0324183
17	6.6783421	4.2195077	5.288267	5.8721178	5.3471075	-.3873815	.1500644
18	6.4982821	4.3944492	5.433722	5.6524892	5.062595	-.7258219	.5268174
19	6.5581978	4.4308168	5.3890717	5.6733233	5.049856	-.6843098	.46828
20	6.8057226	4.2904594	5.164786	6.0591232	5.170484	-.3758329	.1412503
21	7.3158835	4.5951199	5.3936275	5.8916442	5.1357984	-.1211948	.0146882
22	7.3870902	4.6151205	5.1119878	5.811141	5.2574954	.1195903	.0143018

图 11.18 对残差序列进行平方变换后的结果

关于残差序列的意义已在上节论述过，此处不再重复讲解。

图 11.19 是对残差序列的平方值进行对数变换的结果。

	V1	V2	V3	V4	V5	e	ee	lnee
1	4.4067192	.69314718	5.3423343	5.1873858	5.2094862	.1195615	.014295	-4.247848
2	6.4937538	1.0986123	5.32301	5.8607862	5.1590553	1.624628	2.639415	.9705572
3	6.8977049	1.3862944	5.32301	5.8607862	5.1416636	1.81751	3.303343	1.194935
4	5.7525726	1.3862944	5.2094862	5.7745515	5.1119878	.752168	.5657567	-.5695912
5	5.2832037	1.6094379	5.3565863	5.6559918	5.4510385	.182983	.0334828	-3.396724
6	4.5849675	2.1972246	5.3565863	5.6559918	5.2729996	-.9778399	.956171	-.0448185
7	6.8554088	2.3978953	5.288267	5.8721178	5.3278762	1.097735	1.205022	.1864982
8	6.5147127	2.5649494	5.32301	5.8607862	5.0106353	.5566101	.3098148	-1.171781
9	6.2633983	2.5649494	5.3890717	5.6733233	5.0434251	.3636364	.1322315	-2.023201
10	6.2166061	3.0910425	5.1474945	5.0106353	5.236442	.3683498	.1356816	-1.997444
11	7.0850643	3.2188758	5.3423343	5.1873858	5.1357984	.9621836	.9257973	-.0771
12	6.5072777	3.2188758	5.123964	5.9839363	5.1179938	.1360236	.0185024	-3.989853
13	5.8550719	3.5553481	5.198497	5.420535	5.3612922	-.4972979	.2473052	-1.397132
14	6.0473722	3.6635616	5.4380793	5.4638318	5.0998664	-.563445	.3174703	-1.147371
15	6.2166061	3.7612001	5.164786	6.0591232	5.1357984	-.591301	.3496368	-1.05086
16	6.3099183	4.1431347	5.170484	4.634729	5.0814044	-.1800509	.0324183	-3.429032
17	6.6783421	4.2195077	5.288267	5.8721178	5.3471075	-.3873815	.1500644	-1.89669
18	6.4982821	4.3944492	5.433722	5.6524892	5.062595	-.7258219	.5268174	-.6409012
19	6.5581978	4.4308168	5.3890717	5.6733233	5.049856	-.6843098	.46828	-.758689
20	6.8057226	4.2904594	5.164786	6.0591232	5.170484	-.3758329	.1412503	-1.957222
21	7.3158835	4.5951199	5.3936275	5.8916442	5.1357984	-.1211948	.0146882	-4.220713
22	7.3870902	4.6151205	5.1119878	5.811141	5.2574954	.1195903	.0143018	-4.247366

图 11.19　对残差序列的平方值进行对数变换的分析结果

图 11.20 是以上步得到的残差平方对数值为因变量，以 V2 为自变量，并且不包含常数项的最小二乘回归分析结果。

```
. reg lnee V2,nocon

      Source |       SS           df       MS      Number of obs   =       145
-------------+----------------------------------   F(1, 144)       =    448.48
       Model |  2021.97911         1  2021.97911   Prob > F        =    0.0000
    Residual |  649.222688       144  4.50849089   R-squared       =    0.7570
-------------+----------------------------------   Adj R-squared   =    0.7553
       Total |   2671.2018       145  18.4220814   Root MSE        =    2.1233

------------------------------------------------------------------------------
        lnee |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          V2 |  -.5468941   .0258244   -21.18   0.000     -.597938   -.4958502
------------------------------------------------------------------------------
```

图 11.20　最小二乘回归分析结果

图 11.21 是上步进行的最小二乘回归产生的因变量的拟合值结果。

	V1	V2	V3	V4	V5	e	ee	lnee	yhat
1	4.4067192	.69314718	5.3423343	5.1873858	5.2094862	.1195615	.014295	-4.247848	-.3790781
2	6.4937538	1.0986123	5.32301	5.8607862	5.1590553	1.624628	2.639415	.9705572	-.6008245
3	6.8977049	1.3862944	5.32301	5.8607862	5.1416636	1.81751	3.303343	1.194935	-.7581562
4	5.7525726	1.3862944	5.2094862	5.7745515	5.1119878	.752168	.5657567	-.5695912	-.7581562
5	5.2832037	1.6094379	5.3565863	5.6559918	5.4510385	.182983	.0334828	-3.396724	-.880192
6	4.5849675	2.1972246	5.3565863	5.6559918	5.2729996	-.9778399	.956171	-.0448185	-1.201649
7	6.8554088	2.3978953	5.288267	5.8721178	5.3278762	1.097735	1.205022	.1864982	-1.311395
8	6.5147127	2.5649494	5.32301	5.8607862	5.0106353	.5566101	.3098148	-1.171781	-1.402756
9	6.2633983	2.5649494	5.3890717	5.6733233	5.0434251	.3636364	.1322315	-2.023201	-1.402756
10	6.2166061	3.0910425	5.1474945	5.0106353	5.236442	.3683498	.1356816	-1.997444	-1.690473
11	7.0850643	3.2188758	5.3423343	5.1873858	5.1357984	.9621836	.9257973	-.0771	-1.760384
12	6.5072777	3.2188758	5.123964	5.9839363	5.1179938	.1360236	.0185024	-3.989853	-1.760384
13	5.8550719	3.5553481	5.198497	5.420535	5.3612922	-.4972979	.2473052	-1.397132	-1.944399
14	6.0473722	3.6635616	5.4380793	5.4638318	5.0998664	-.563445	.3174703	-1.147371	-2.00358
15	6.2166061	3.7612001	5.164786	6.0591232	5.1357984	-.591301	.3496368	-1.05086	-2.056978
16	6.3099183	4.1431347	5.170484	4.634729	5.0814044	-.1800509	.0324183	-3.429032	-2.265856
17	6.6783421	4.2195077	5.288267	5.8721178	5.3471075	-.3873815	.1500644	-1.89669	-2.307624
18	6.4982821	4.3944492	5.433722	5.6524892	5.062595	-.7258219	.5268174	-.6409012	-2.403298
19	6.5581978	4.4308168	5.3890717	5.6733233	5.049856	-.6843098	.46828	-.758689	-2.423187
20	6.8057226	4.2904594	5.164786	6.0591232	5.170484	-.3758329	.1412503	-1.957222	-2.346427
21	7.3158835	4.5951199	5.3936275	5.8916442	5.1357984	-.1211948	.0146882	-4.220713	-2.513044
22	7.3870902	4.6151205	5.1119878	5.811141	5.2574954	.1195903	.0143018	-4.247366	-2.523982

图 11.21　最小二乘回归分析产生的因变量的拟合值结果

图 11.22 是对因变量的拟合值进行指数变换的结果。

	V1	V2	V3	V4	V5	e	ee	lnee	yhat	yhathat
1	4.4067192	.69314718	5.3423343	5.1873858	5.2094862	.1195615	.014295	-4.247848	-.3790781	.6844922
2	6.4937538	1.0986123	5.32301	5.8607862	5.1590553	1.624628	2.639415	.9705572	-.6008245	.5483593
3	6.8977049	1.3862944	5.32301	5.8607862	5.1416636	1.81751	3.303343	1.194935	-.7581562	.4685295
4	5.7525726	1.3862944	5.2094862	5.7745515	5.1119878	.752168	.5657567	-.5695912	-.7581562	.4685295
5	5.2832037	1.6094379	5.3565863	5.6559918	5.4510385	.182983	.0334828	-3.396724	-.880192	.4147032
6	4.5849675	2.1972246	5.3565863	5.6559918	5.2729996	-.9778399	.956171	-.0448185	-1.201649	.3006979
7	6.8554088	2.3978953	5.288267	5.8721178	5.3278762	1.097735	1.205022	.1864982	-1.311395	.269444
8	6.5147127	2.5649494	5.32301	5.8607862	5.0106353	.5566101	.3098148	-1.171781	-1.402756	.2459184
9	6.2633983	2.5649494	5.3890717	5.6733233	5.0434251	.3636364	.1322315	-2.023201	-1.402756	.2459184
10	6.2166061	3.0910425	5.1474945	5.0106353	5.236442	.3683498	.1356816	-1.997444	-1.690473	.1844323
11	7.0850643	3.2188758	5.3423343	5.1873858	5.1357984	.9621836	.9257973	-.0771	-1.760384	.1719788
12	6.5072777	3.2188758	5.123964	5.9839363	5.1179938	.1360236	.0185024	-3.989853	-1.760384	.1719788
13	5.8550719	3.5553481	5.198497	5.420535	5.3612922	-.4972979	.2473052	-1.397132	-1.944399	.1430732
14	6.0473722	3.6635616	5.4380793	5.4638318	5.0998664	-.563445	.3174703	-1.147371	-2.00358	.1348516
15	6.2166061	3.7612001	5.164786	6.0591232	5.1357984	-.591301	.3496368	-1.05086	-2.056978	.1278397
16	6.3099183	4.1431347	5.170484	4.634729	5.0814044	-.1800509	.0324183	-3.429032	-2.265856	.1037412
17	6.6783421	4.2195077	5.288267	5.8721178	5.3471075	-.3873815	.1500644	-1.89669	-2.307624	.0994974
18	6.4982821	4.3944492	5.433722	5.6524892	5.062595	-.7258219	.5268174	-.6409012	-2.403298	.0904192
19	6.5581978	4.4308168	5.3890717	5.6733233	5.049856	-.6843098	.46828	-.758689	-2.423187	.0886386
20	6.8057226	4.2904594	5.164786	6.0591232	5.170484	-.3758329	.1412503	-1.957222	-2.346427	.0957105
21	7.3158835	4.5951199	5.3936275	5.8916442	5.1357984	-.1211948	.0146882	-4.220713	-2.513044	.0810212
22	7.3870902	4.6151205	5.1119878	5.811141	5.2574954	.1195903	.0143018	-4.247366	-2.523982	.0801399
23	7.0273145	4.7791235	5.2574954	5.4161004	5.0998664	-.2879924	.0829396	-2.489643	-2.613674	.0732648

图 11.22 对因变量的拟合值进行指数变换的结果

图 11.23 是加权最小二乘回归分析的结果。

```
. reg  V1 V2 V3 V4 V5 [aw=1/yhathat]
(sum of wgt is   7.8139e+03)

      Source |       SS           df       MS      Number of obs   =       145
-------------+----------------------------------   F(  4,    140)  =    888.79
       Model |  173.679487         4  43.4198717   Prob > F        =    0.0000
    Residual |  6.83940919       140  .048852923   R-squared       =    0.9621
-------------+----------------------------------   Adj R-squared   =    0.9610
       Total |  180.518896       144  1.25360344   Root MSE        =    .22103

------------------------------------------------------------------------------
          V1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          V2 |   .8733637   .0155164    56.29   0.000     .8426869    .9040405
          V3 |   .5411784   .1713419     3.16   0.002     .2024263    .8799305
          V4 |   .4642838   .0625673     7.42   0.000     .3405849    .5879827
          V5 |  -.0882917   .1977227    -0.45   0.656    -.4792002    .3026168
       _cons |  -2.157215   1.376615    -1.57   0.119    -4.878857    .5644262
------------------------------------------------------------------------------
```

图 11.23 加权最小二乘回归分析的结果

在上面的分析结果中看出模型的 F 值（代表模型的显著程度）、部分变量的 P 值以及 R-squared 值、Adj R-squared 值（代表模型的解释能力）都较普通最小二乘回归分析有了一定程度的优化，这就是克服异方差带来的改善效果。

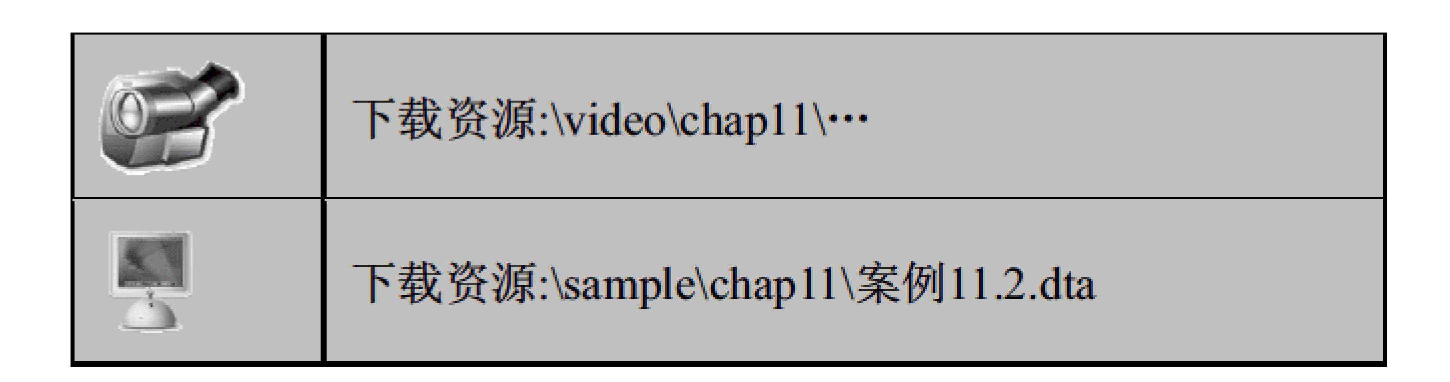
11.2 实例二——自相关检验与应对

11.2.1 自相关检验与应对的功能与意义

如果线性相关模型中的随机误差项的各期望值之间存在着相关关系，这时，我们就称随机误差项之间存在自相关性（Autocorrelation）。线性回归模型中随机误差项存在序列相关的原因很多，但主要是由经济变量自身特点、数据特点、变量选择及模型函数的形式选择引起的。常见原因包括经济变量惯性的作用、经济行为的滞后性、一些随机因素的干扰或影响、模型设定误差、观测数据处理等。自相关不会影响到最小二乘估计量的线性和无偏性，但会使之失去

有效性，使之不再是最优估计量，而且自相关的系数估计量将有相当大的方差，T 检验也不再显著，模型的预测功能失效，所以在进行回归分析时往往需要检验数据的自相关性，从而提出针对性的解决方案。常用的用于判断数据是否存在自相关的检验方法有绘制残差序列图、BG 检验、Box-Pierce Q 检验、DW 检验等，解决自相关的方法有使用自相关异方差稳健的标准差进行回归以及使用广义最小二乘回归分析方法进行回归等。

11.2.2 相关数据来源



【例 11.2】表 11.2 给出了某企业经营利润和经营资产的有关数据，试使用经营利润作为因变量，以经营资产作为自变量，对这些数据使用最小二乘回归分析的方法进行研究，并进行自相关检验，最终建立合适的回归方程模型用于描述变量之间的关系。

表 11.2 某企业经营利润和经营资产的有关数据

月份	经营利润/万元	经营资产/万元
1	22.89	283.9
2	23.15	286.9
3	24.12	291.5
4	25.19	303.33
5	27.02	314.49
6	25.52	310.25
…	…	…
45	66.32	456.05
46	63.12	470.3
47	59.89	472.69
48	58.49	512.9
49	67.79	550.96

11.2.3 Stata 分析过程

在利用 Stata 进行分析之前，要把数据录入到 Stata 中。本例中有 3 个变量，分别是月份、经营利润和经营资产。把月份变量设定为 month，把经营利润变量设定为 profit，把经营资产变量设定为 asset，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作已在第 1 章中有过详细讲述。录入完成后数据如图 11.24 所示。

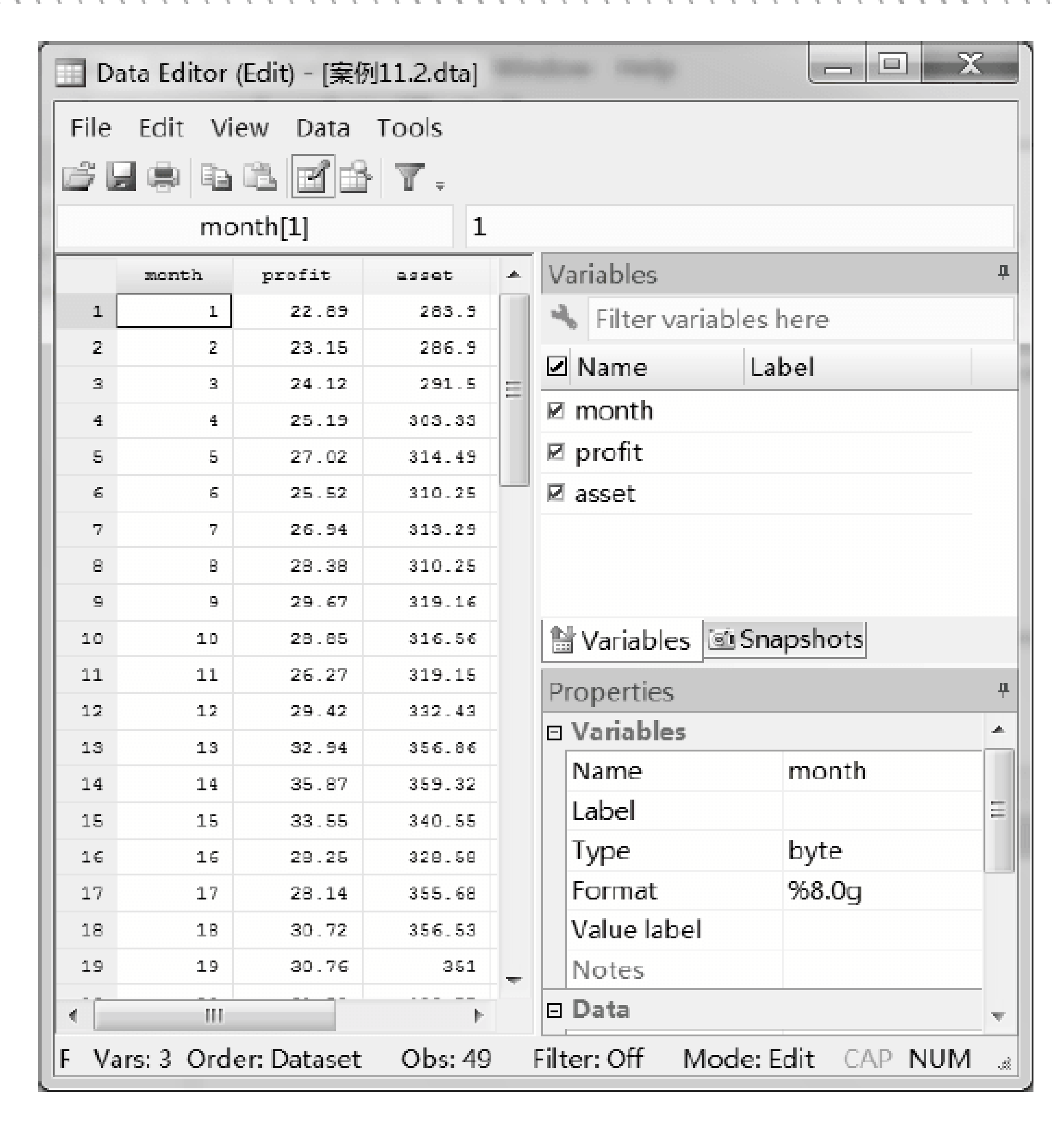


图 11.24　案例 11.2 数据

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入如下命令。

- summarize month profit asset,detail：本命令旨在对数据进行描述性分析，从总体上探索数据特征，观测其是否存在极端数据或者变量间的量纲差距过大，从而可能会对回归分析结果造成不利影响。
- correlate month profit asset：本命令旨在对数据进行相关性分析，旨在探索变量之间尤其是因变量与各个自变量之间的相关性关系，该步骤是进行回归分析前的必要准备。
- regress profit asset：本命令旨在对数据进行回归分析，用于探索自变量对因变量的影响情况。
- vce：本命令旨在获得变量的方差-协方差矩阵。
- test asset：本命令旨在检验回归分析获得的各个自变量系数的显著性。
- predict yhat：本命令旨在获得因变量的拟合值。
- predict e,resid：本命令旨在获得回归模型的估计残差。
- tsset month：本命令旨在把数据定义为以 month 为周期的时间序列。
- scatter e l.e：本命令旨在绘制残差与残差滞后一期的散点图，用于探索数据是否存在一阶自相关。
- ac e：本命令旨在绘制残差的自相关图，用于探索其自相关阶数。
- pac e：本命令旨在绘制残差的偏自相关图，用于探索其自相关阶数。
- estat bgodfrey：本命令为 BG 检验，旨在检验残差自相关性。
- wntestq e：本命令为 Box-Pierce Q 检验，旨在检验残差自相关性。
- estat dwatson：本命令为 DW 检验，旨在检验残差自相关性。

- di 49^0.25：本命令为计算样本个数的 1/4 次幂，旨在确定使用异方差自相关稳健的标准差进行回归的滞后阶数。
- newey profit asset,lag(3)：本命令为采用异方差自相关稳健的标准差对数据进行回归分析，克服数据的自相关性对最小二乘回归分析造成的不利影响。

03 设置完毕后，按键盘上的回车键，等待输出结果。

11.2.4 结果分析

在 Stata 14.0 主界面的结果窗口可以看到如图 11.25~图 11.42 所示的分析结果。

1．对数据进行描述性分析的结果

图 11.25 是对数据进行描述性分析的结果。关于这一分析过程对于回归分析的重要意义已在前面章节中论述过，此处不再重复讲解。

```
. summarize month profit asset,detail

                           month
-------------------------------------------------------------
      Percentiles      Smallest
 1%            1              1
 5%            3              2
10%            5              3       Obs                  49
25%           13              4       Sum of Wgt.          49

50%           25                      Mean                 25
                        Largest       Std. Dev.      14.28869
75%           37             46
90%           45             47       Variance       204.1667
95%           47             48       Skewness              0
99%           49             49       Kurtosis          1.799

                           profit
-------------------------------------------------------------
      Percentiles      Smallest
 1%        22.89          22.89
 5%        24.12          23.15
10%        25.52          24.12       Obs                  49
25%        28.85          25.19       Sum of Wgt.          49

50%        34.74                      Mean           39.50796
                        Largest       Std. Dev.      13.07854
75%        48.46          63.12
90%        59.89          64.97       Variance       171.0482
95%        64.97          66.32       Skewness       .6806106
99%        67.79          67.79       Kurtosis       2.213728

                           asset
-------------------------------------------------------------
      Percentiles      Smallest
 1%        283.9          283.9
 5%        291.5          286.9
10%       310.25          291.5       Obs                  49
25%       332.43         303.33       Sum of Wgt.          49

50%       391.99                      Mean           385.0224
                        Largest       Std. Dev.      60.03378
75%       424.15          470.3
90%       456.05         472.69       Variance       3604.055
95%       472.69          512.9       Skewness       .3029836
99%       550.96         550.96       Kurtosis        2.83925
```

图 11.25　描述性分析的结果

在如图 11.25 所示的分析结果中，可以得到很多信息，包括百分位数、4 个最小值、4 个最大值、平均值、标准差、偏度、峰度等。

（1）百分位数（Percentiles）

可以看出变量 month 的第 1 个四分位数（25%）是 13，第 2 个四分位数（50%）是 25，第 3 个四分位数（75%）是 37；变量 profit 的第 1 个四分位数（25%）是 28.85，第 2 个四分

位数（50%）是 34.74，第 3 个四分位数（75%）是 48.46；变量 asset 的第 1 个四分位数（25%）是 332.43，第 2 个四分位数（50%）是 391.99，第 3 个四分位数（75%）是 424.15。

（2）4 个最小值（Smallest）

变量 month 最小的 4 个数据值分别是 1、2、3、4

变量 profit 最小的 4 个数据值分别是 22.89、23.15、24.12、25.19。

变量 asset 最小的 4 个数据值分别是 283.9、286.9、291.5、303.33。

（3）4 个最大值（Largest）

变量 month 最大的 4 个数据值分别是 46、47、48、49。

变量 profit 最大的 4 个数据值分别是 63.12、64.97、66.32、67.79。

变量 asset 最大的 4 个数据值分别是 470.3、472.69、512.9、550.96。

（4）平均值（Mean）和标准差（Std. Dev）

变量 month 的平均值为 25，标准差是 14.28869。

变量 profit 的平均值为 39.50796，标准差是 13.07854。

变量 asset 的平均值为 385.0224，标准差是 60.03378。

（5）偏度（Skewness）和峰度（Kurtosis）

变量 month 的偏度为 0，为无偏度。

变量 profit 的偏度为 0.6806106，为正偏度但不大。

变量 asset 的偏度为 0.3029836，为正偏度但不大。

变量 month 的峰度为 1.799，有一个比正态分布略短的尾巴。

变量 profit 的峰度为 2.213728，有一个比正态分布略短的尾巴。

变量 asset 的峰度为 2.83925，有一个比正态分布略短的尾巴。

综上所述，数据的总体质量还是可以的，没有极端异常值，变量间的量纲差距、变量的偏度、峰度也是可以接受的，可以进入下一步的分析。

2．对数据进行相关性分析的结果

图 11.26 是对数据进行相关性分析的结果。关于这一分析过程对于回归分析的重要意义已在前面章节中论述过，此处不再重复讲解。

```
. correlate month profit asset
(obs=49)

             |    month   profit    asset
-------------+---------------------------
       month |   1.0000
      profit |   0.9377   1.0000
       asset |   0.9557   0.8917   1.0000
```

图 11.26　相关性分析的结果

在图 11.26 中，profit 与 asset 之间的相关关系还是可以接受的，可以进入下面的回归分析过程。

3．对数据进行回归分析的结果

图 11.27 是对数据进行回归分析的结果。

```
. regress   profit asset

      Source |       SS       df       MS              Number of obs =      49
-------------+------------------------------           F(  1,    47) =  182.40
       Model |  6528.14552     1  6528.14552           Prob > F      =  0.0000
    Residual |  1682.16623    47  35.7907709           R-squared     =  0.7951
-------------+------------------------------           Adj R-squared =  0.7908
       Total |  8210.31175    48  171.048161           Root MSE      =  5.9825

------------------------------------------------------------------------------
      profit |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       asset |   .1942579   .0143837    13.51   0.000     .1653217     .223194
       _cons |  -35.28568   5.603588    -6.30   0.000    -46.55864   -24.01271
------------------------------------------------------------------------------
```

图 11.27　回归分析的结果

从上述分析结果中，可以看出共有 49 个样本参与了分析，模型的 F 值(1, 47) = 182.40，P 值（Prob > F） = 0.0000，说明模型整体上是非常显著的。模型的可决系数（R-squared）=0.7951，模型修正的可决系数（Adj R-squared） = 0.7908，说明模型的解释能力非常不错。

模型的回归方程是：

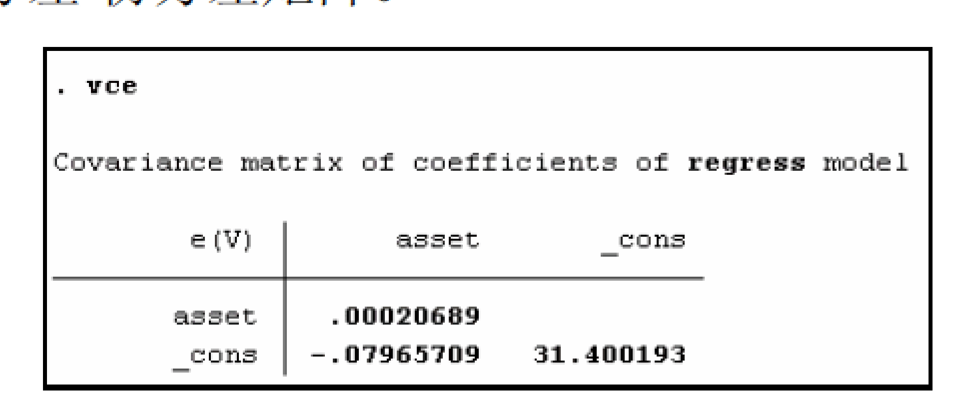
profit =0.1942579* asset -35.28568

变量 asset 的系数标准误是 0.0143837，t 值为 13.51，P 值为 0.000，系数是非常显著的，95%的置信区间为[0.1653217, 0.223194]。常数项的系数标准误是 5.603588，t 值为-6.30，P 值为 0.000，系数也是非常不显著的，95%的置信区间为[-46.55864, -24.01271]。

从上面的分析可以看出该企业的经营利润与经营资产之间是一种正向联动变化关系，但是经营资产的增加仅能带来经营利润近五分之一的增加。

4．变量的方差-协方差矩阵结果

图 11.28 是对变量的方差-协方差矩阵。



```
. vce

Covariance matrix of coefficients of regress model

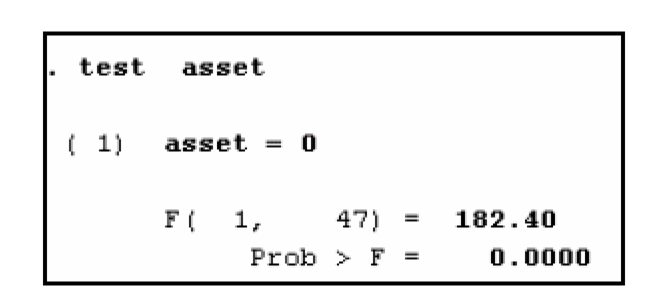
        e(V) |      asset       _cons
-------------+------------------------
       asset |  .00020689
       _cons | -.07965709   31.400193
```

图 11.28　变量的方差-协方差矩阵

从图 11.28 中可以看出，变量与常数项系数的方差与协方差都不是很大。

5．对变量系数的假设检验结果

图 11.29 是对变量系数的假设检验结果。



```
. test   asset

 ( 1)   asset = 0

       F(  1,    47) =  182.40
            Prob > F =    0.0000
```

图 11.29　对变量系数的假设检验结果

从图 11.29 中可以看出，模型非常显著，在 5%的显著性水平上通过了检验。

6．对因变量的拟合值的预测

图 11.30 是对因变量的拟合值的预测。

	month	profit	asset	yhat
1	1	22.89	283.9	19.86413
2	2	23.15	286.9	20.4469
3	3	24.12	291.5	21.34049
4	4	25.19	303.33	23.63856
5	5	27.02	314.49	25.80647
6	6	25.52	310.25	24.98282
7	7	26.94	313.29	25.57337
8	8	28.38	310.25	24.98282
9	9	29.67	319.16	26.71366
10	10	28.85	316.56	26.20859
11	11	26.27	319.15	26.71172
12	12	29.42	332.43	29.29146
13	13	32.94	356.86	34.03718
14	14	35.87	359.32	34.51506
15	15	33.55	340.55	30.86884
16	16	28.25	328.58	28.54357
17	17	28.14	355.68	33.80796
18	18	30.72	356.53	33.97308
19	19	30.76	351	32.89883
20	20	31.59	359.77	34.60247
21	21	28.29	378.95	38.32834
22	22	30.33	378.92	38.32251
23	23	31.09	389.91	40.45741
24	24	32.15	407.59	43.89188
25	25	34.27	413.99	45.13513

图 11.30　对因变量的拟合值的预测

因变量预测拟合值是根据自变量的值和得到的回归方程计算出来的，主要用于预测未来。在图 11.30 中可以看到 yhat 的值与 profit 的值是比较相近的，所以拟合的回归模型还是不错的。

7．回归分析得到的残差序列

图 11.31 是回归分析得到的残差序列。

	month	profit	asset	yhat	e
1	1	22.89	283.9	19.86413	3.025871
2	2	23.15	286.9	20.4469	2.703098
3	3	24.12	291.5	21.34049	2.779512
4	4	25.19	303.33	23.63856	1.551444
5	5	27.02	314.49	25.80647	1.213525
6	6	25.52	310.25	24.98282	.5371766
7	7	26.94	313.29	25.57337	1.366631
8	8	28.38	310.25	24.98282	3.397175
9	9	29.67	319.16	26.71366	2.956338
10	10	28.85	316.56	26.20859	2.64141
11	11	26.27	319.15	26.71172	-.4417171
12	12	29.42	332.43	29.29146	.1285384
13	13	32.94	356.86	34.03718	-1.097181
14	14	35.87	359.32	34.51506	1.354941
15	15	33.55	340.55	30.86884	2.681165
16	16	28.25	328.58	28.54357	-.2935677
17	17	28.14	355.68	33.80796	-5.667957
18	18	30.72	356.53	33.97308	-3.253078
19	19	30.76	351	32.89883	-2.138831
20	20	31.59	359.77	34.60247	-3.012471
21	21	28.29	378.95	38.32834	-10.03834
22	22	30.33	378.92	38.32251	-7.992514
23	23	31.09	389.91	40.45741	-9.367405
24	24	32.15	407.59	43.89188	-11.74188
25	25	34.27	413.99	45.13513	-10.86513
26	26	34.74	405.33	43.45286	-8.712856

图 11.31　残差序列

关于残差序列的意义已在上节中论述过，此处不再重复讲解。

8．以 month 为周期的时间序列的结果

图 11.32 是把数据定义成以 month 为周期的时间序列的结果。



```
. tsset month
        time variable:  month, 1 to 49
                delta:  1 unit
```

图 11.32　以 month 为周期的时间序列的结果

关于时间序列的相关概念与分析方法等，将在后续的章节中详细进行说明，这里不再赘述。

9．散点图

图 11.33 是残差与残差滞后一期的散点图。

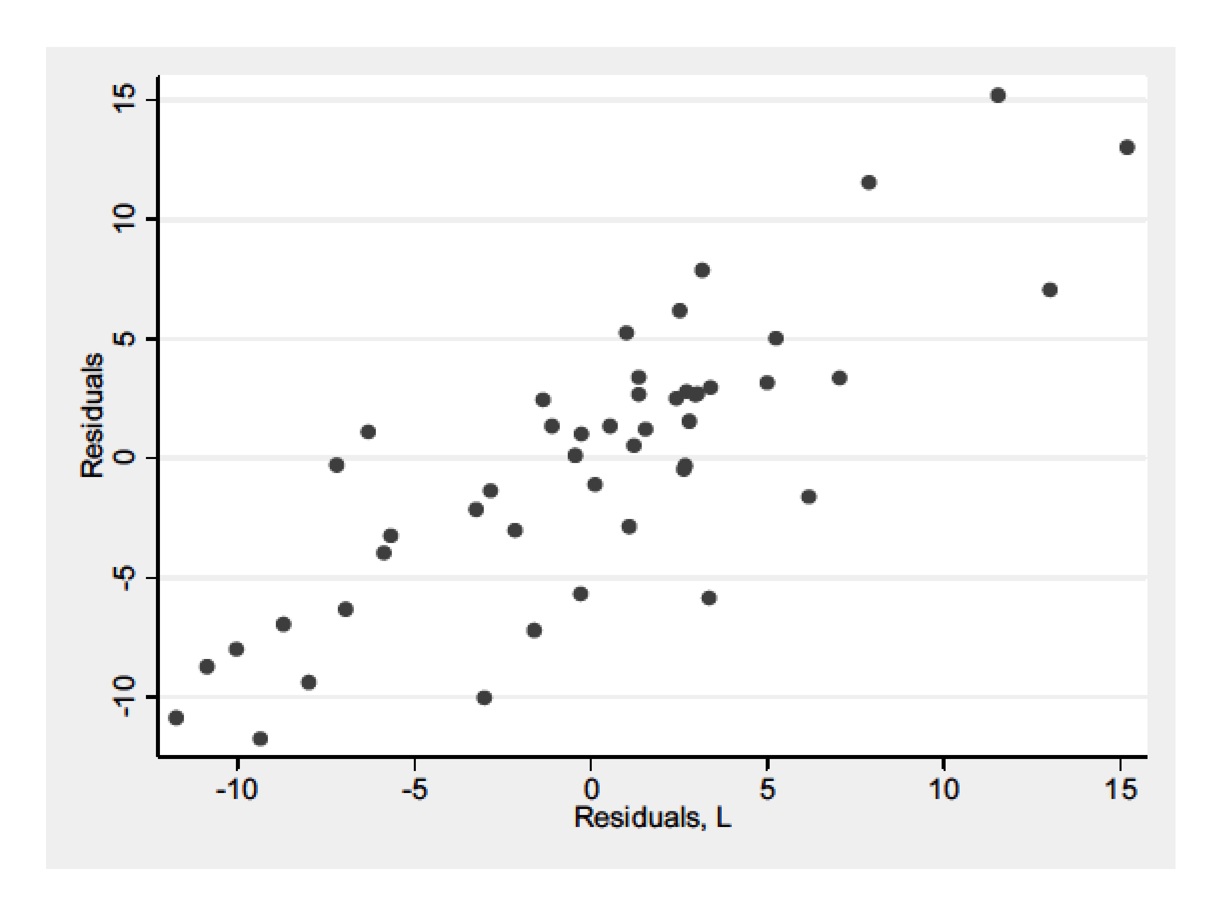


图 11.33　残差与残差滞后一期的散点图

从图 11.33 中可以看出，残差与滞后一期的残差之间存在着一种类似正向线性变动关系，所以数据是存在自相关的。

10．自相关图

图 11.34 是残差序列的自相关图。

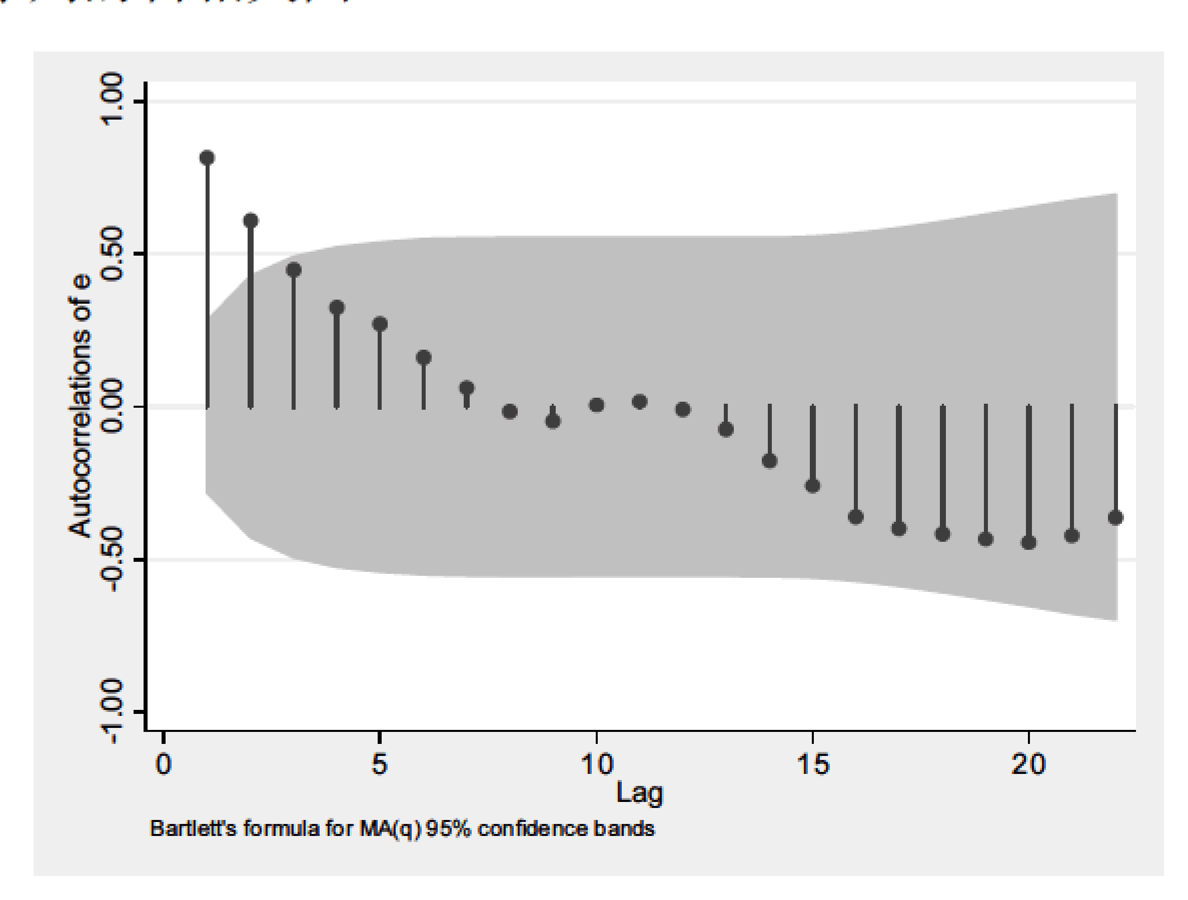


图 11.34　残差序列的自相关图

图 11.34 中的横轴表示滞后阶数，阴影部分表示 95%的自相关置信区间，在阴影部分之外表示自相关系数显著不为 0，从图 11.34 中可以看出，数据主要是存在一阶自相关的。

11．偏自相关图

图 11.35 是残差序列的偏自相关图。

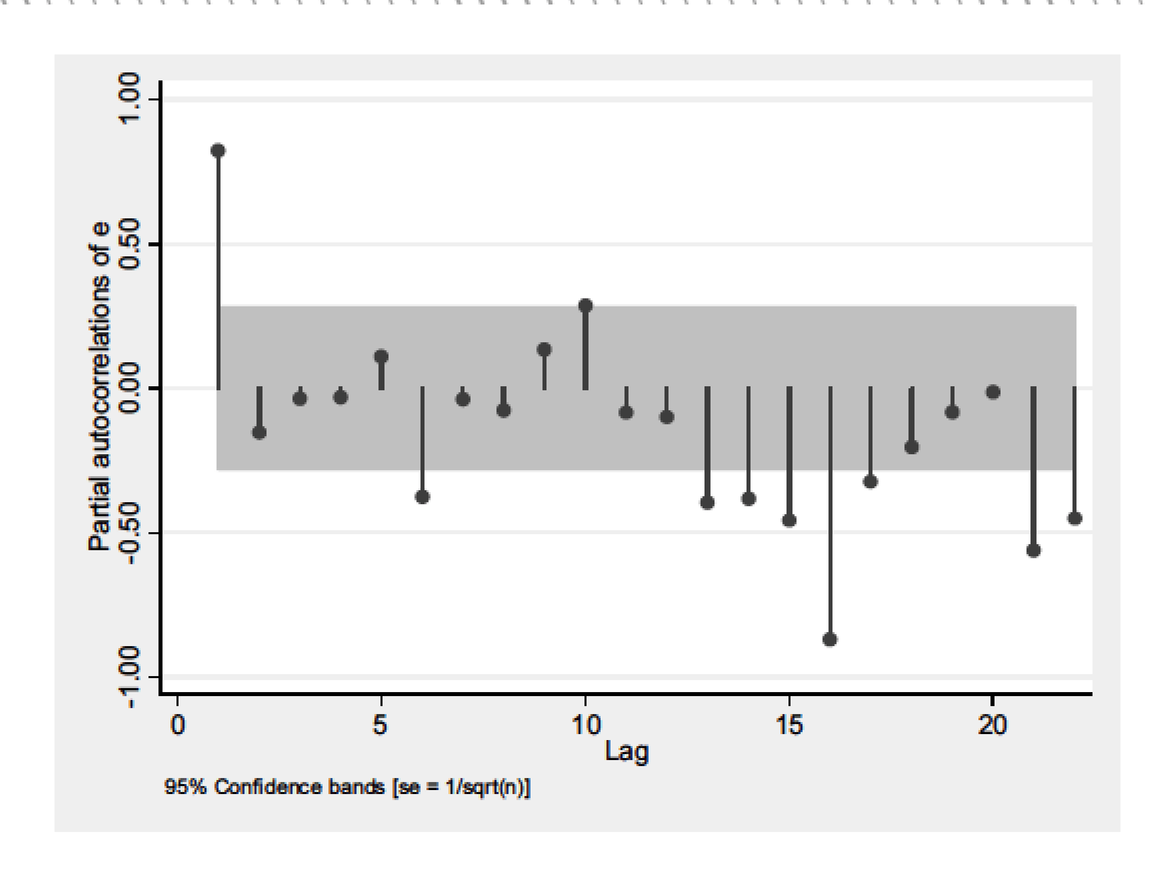


图 11.35　残差序列的偏自相关图

图 11.35 中的横轴表示滞后阶数，阴影部分表示 95%的自相关置信区间，在阴影部分之外表示自相关系数显著不为 0，从图 11.35 中同样可以看出，数据主要是存在一阶自相关的。

12．BG 检验的检验结果

图 11.36 是 BG 检验的检验结果。

```
. estat bgodfrey

Breusch-Godfrey LM test for autocorrelation
```

lags(p)	chi2	df	Prob > chi2
1	33.069	1	0.0000

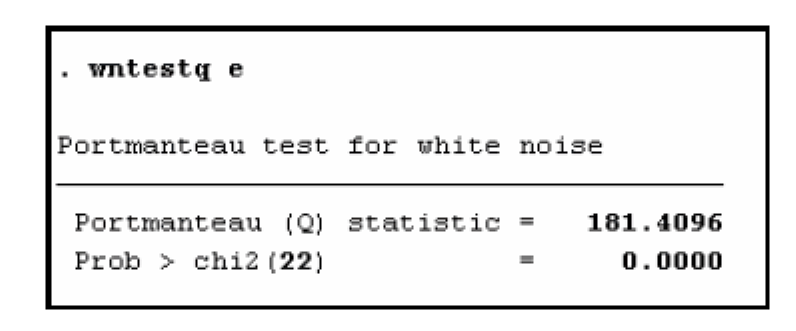
```
H0: no serial correlation
```

图 11.36　BG 检验的检验结果

BG 检验的原假设是数据没有自相关。从图 11.36 中可以看出，P 值为 0.0000，非常显著地拒绝了无自相关的原假设，认为存在自相关。

13．Box-Pierce Q 检验的检验结果

图 11.37 是 Box-Pierce Q 检验的检验结果。



```
. wntestq e

Portmanteau test for white noise

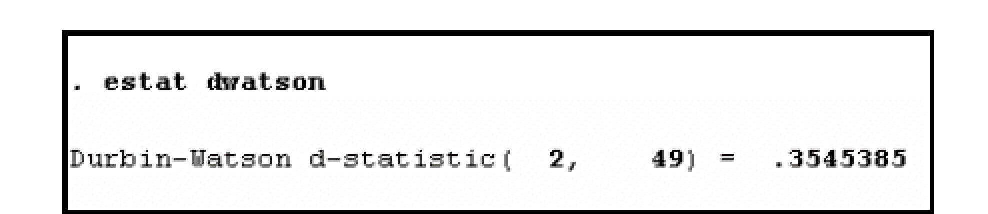
 Portmanteau (Q) statistic =   181.4096
 Prob > chi2(22)           =     0.0000
```

图 11.37　Box-Pierce Q 检验的检验结果

Box-Pierce Q 检验的原假设是数据没有自相关。从图 11.37 中可以看出，P 值为 0.0000，非常显著地拒绝了无自相关的原假设，认为存在自相关。

14．DW 检验的检验结果

图 11.38 是 DW 检验的检验结果。



```
. estat dwatson

Durbin-Watson d-statistic(  2,    49) =  .3545385
```

图 11.38　DW 检验的检验结果

DW 检验的原假设数据没有自相关。从图 11.38 中可以看出，DW 值为 0.3545385，远远小于无自相关时的值 2，所以认为存在正的自相关。

图 11.39 是计算样本个数的 1/4 次幂的结果。



```
. di 49^0.25
2.6457513
```

图 11.39　计算样本个数的 1/4 次幂的结果

本例中，样本个数为 49，49 的 0.25 次方是 2.6457513，所以确定的滞后阶数是 3。

图 11.40 是使用自相关异方差稳健的标准差对数据进行回归分析的结果。

```
. newey profit asset,lag(3)

Regression with Newey-West standard errors          Number of obs  =        49
maximum lag: 3                                      F(  1,     47)  =    107.43
                                                    Prob > F        =    0.0000

------------------------------------------------------------------------------
             |             Newey-West
      profit |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       asset |   .1942579   .0187418    10.36   0.000     .1565543    .2319615
       _cons |  -35.28568   6.344974    -5.56   0.000    -48.05012   -22.52123
------------------------------------------------------------------------------
```

图 11.40　使用自相关异方差稳健的标准差对数据进行回归分析的结果

从上述分析结果中可以看出，模型整体的显著性、自变量与常数项系数的显著性以及模型的解释能力依旧很高。

11.2.5　案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

下面使用广义最小二乘回归分析方法解决数据的异方差问题。

以本例为例进行说明，操作命令如下。

- prais profit asset,corc：本命令旨在对数据进行以 profit 为因变量、以 asset 为自变量的迭代式 CO 估计法广义最小二乘回归分析。
- prais profit asset,nolog：本命令旨在对数据进行以 profit 为因变量、以 asset 为自变量的迭代式 PW 估计法广义最小二乘回归分析。

在命令窗口输入命令并按回车键进行确认，结果如图 11.41~图 11.42 所示。

图 11.41 是对数据进行迭代式 CO 估计法广义最小二乘回归分析的结果。

```
Cochrane-Orcutt AR(1) regression -- iterated estimates

      Source |       SS       df       MS              Number of obs =      48
-------------+------------------------------           F(  1,    46) =    3.94
       Model |  38.9070104     1  38.9070104           Prob > F      =  0.0531
    Residual |  453.948232    46  9.86843982           R-squared     =  0.0789
-------------+------------------------------           Adj R-squared =  0.0589
       Total |  492.855242    47  10.4862817           Root MSE      =  3.1414

------------------------------------------------------------------------------
      profit |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       asset |    .069753   .0351296     1.99   0.053    -.0009592    .1404652
       _cons |   29.04086   23.03048     1.22   0.229     -18.9274    77.00912
-------------+----------------------------------------------------------------
         rho |   .9672991
------------------------------------------------------------------------------
Durbin-Watson statistic (original)    0.354538
Durbin-Watson statistic (transformed) 1.927109
```

图 11.41　对数据进行迭代式 CO 估计法广义最小二乘回归分析的结果

对本结果的详细解读与前面类似，此处限于篇幅不再赘述。但值得注意的是 DW 值从 0.354538 跃升至 1.927109，非常接近于没有自相关时的值 2，所以经过 CO 迭代变换后，模型消除了自相关，但是模型的显著程度和解释能力都有所下降，这也是必须付出的代价。

图 11.42 是对数据进行迭代式 PW 估计法广义最小二乘回归分析的结果。

```
. prais profit asset,nolog


Prais-Winsten AR(1) regression -- iterated estimates

      Source |       SS       df       MS              Number of obs =      49
-------------+------------------------------           F(  1,    47) =    7.55
       Model |  75.5863133     1  75.5863133           Prob > F      =  0.0085
    Residual |  470.661312    47  10.0140705           R-squared     =  0.1384
-------------+------------------------------           Adj R-squared =  0.1200
       Total |  546.247626    48  11.3801589           Root MSE      =  3.1645

------------------------------------------------------------------------------
      profit |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       asset |   .1046879    .029304     3.57   0.001      .045736    .1636399
       _cons |   .0516432   12.70555     0.00   0.997    -25.50864    25.61192
-------------+----------------------------------------------------------------
         rho |   .9291977
------------------------------------------------------------------------------
Durbin-Watson statistic (original)    0.354538
Durbin-Watson statistic (transformed) 1.861233
```

图 11.42　对数据进行迭代式 PW 估计法广义最小二乘回归分析的结果

对本结果的详细解读与前面类似，此处限于篇幅不再赘述。但值得注意的是 DW 值从 0.354538 跃升至 1.861233，非常接近于没有自相关时的值 2，所以经过 PW 迭代变换后，模型消除了自相关，同样，模型的显著程度和解释能力也有所下降。

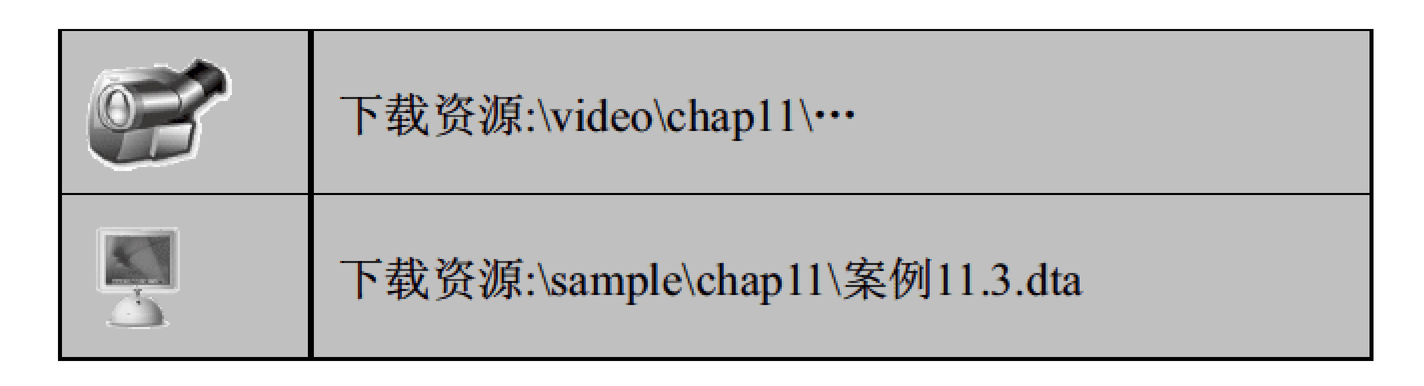
11.3　实例三——多重共线性检验与应对

11.3.1　多重共线性检验与应对的功能与意义

多重共线性包括严重的多重共线性和近似的多重共线性。在进行回归分析时，如果某一自变量可以被其他的自变量通过线性组合得到，那么数据就存在严重的多重共线性问题。近似的多重共线性是指某自变量能够被其他的自变量较多地解释，或者说自变量之间存在着很大程度的信息重叠。在数据存在多重共线性的情况下，最小二乘回归分析得到的系数值仍然是最优无偏估计的，但是会导致系数的估计值不准确，而且会使部分系数的显著性很弱，也不好区分

每个自变量对因变量的影响程度。解决多重共线性的办法通常有两种：一种是剔除不显著的变量；另外一种是进行因子分析提取出相关性较弱的几个主因子再进行回归分析。

11.3.2　相关数据来源



【例 11.3】表 11.3 给出了我国 1996－2003 年国民经济主要指标统计数据。试使用国内生产总值作为因变量，以货物周转量、原煤、发电量、原油等作为自变量，对这些数据使用最小二乘回归分析的方法进行研究，并进行多重共线性检验，最终建立合适的回归方程模型用于描述变量之间的关系。

表 11.3　我国 1996－2003 年国民经济主要指标统计数据

年份	国内生产总值/亿元	货物周转量/亿吨千米	原煤/亿吨	发电量/亿千瓦时	原油/万吨
1996	67 884.6	36 590.0	14.0	10 813.0	15 733.0
1997	74 462.6	38 385.0	13.7	11 356.0	16 074.0
1998	78 345.0	38 089.0	12.5	11 670.0	16 100.0
1999	82 067.0	40 568.0	10.5	12 393.0	16 000.0
2000	89 442.0	44 321.0	10.0	13 556.0	16 300.0
2001	97 315.0	47 710.0	11.6	14 808.0	16 396.0
2002	105 172.0	50 686.0	13.8	16 540.0	16 700.0
2003	117 251.9	53 859.0	16.7	19 106.0	16 960.0

11.3.3　Stata 分析过程

在用 Stata 进行分析之前，要把数据录入到 Stata 中。本例中有 6 个变量，分别是年份、国内生产总值、货物周转量、原煤、发电量、原油。我们把这 6 个变量分别设定为 V1、V2、V3、V4、V5、V6，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作在第 1 章中已有详细讲述。录入完成后数据如图 11.43 所示。

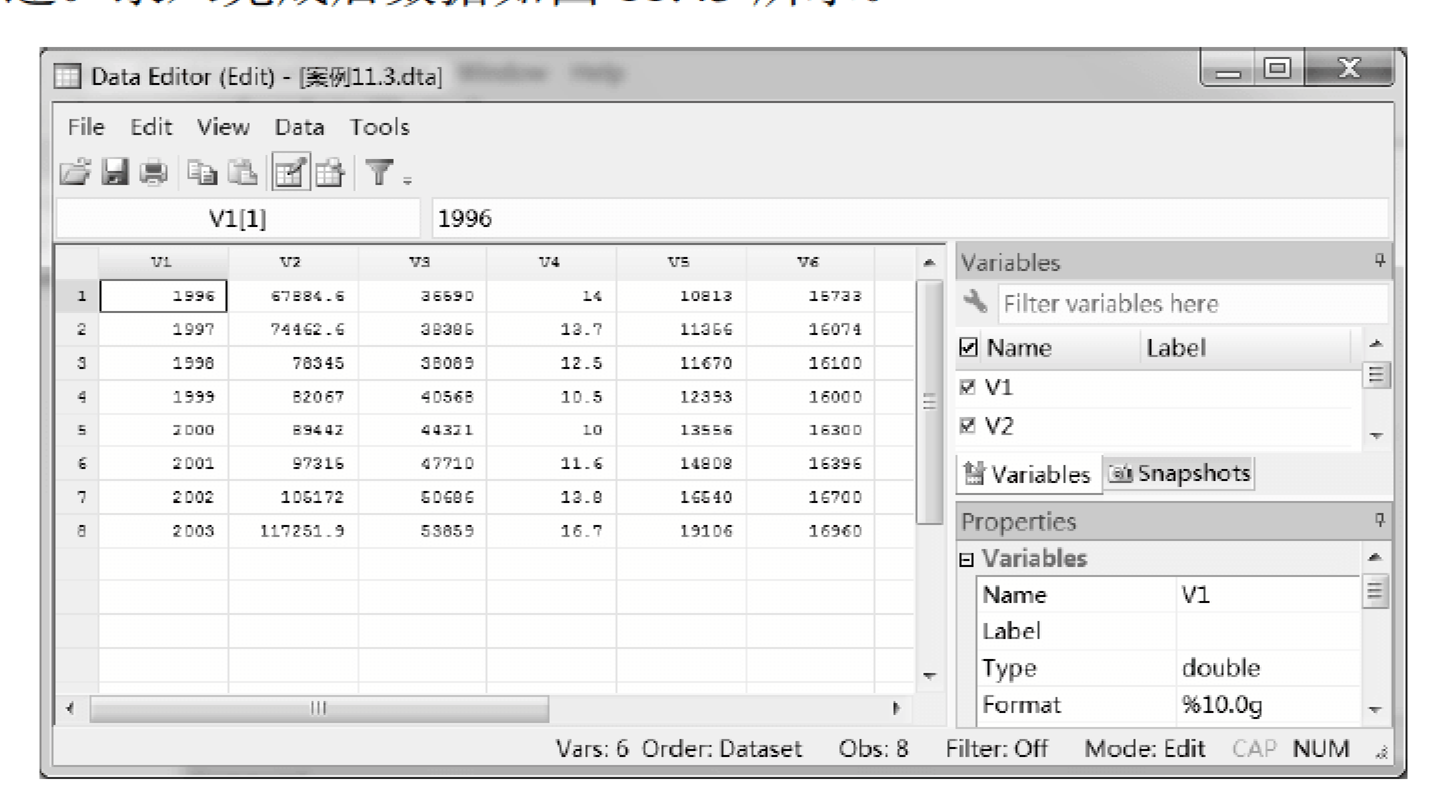


图 11.43　案例 11.3 数据

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入如下命令。

- summarize　V1 V2 V3 V4 V5 V6,detail：本命令旨在对数据进行描述性分析，从总体上探索数据特征，观测其是否存在极端数据或者变量间的量纲差距过大，从而可能会对回归分析结果造成不利影响。
- correlate　V1 V2 V3 V4 V5 V6：本命令旨在对数据进行相关性分析，旨在探索变量之间尤其是因变量与各个自变量之间的相关性关系，该步骤是进行回归分析前的必要准备。
- regress　V2 V3 V4 V5 V6：本命令旨在对数据进行回归分析，探索自变量对因变量的影响情况。
- estat vif：本命令旨在对模型进行多重共线性检验。
- regress　V2 V3 V4 V6：本命令旨在上步的基础上剔除最大的方差膨胀因子然后再重新进行回归。
- estat vif：本命令旨在对新模型进行多重共线性检验。
- regress　V2 V3 V4：本命令旨在上步的基础上剔除最大的方差膨胀因子，然后再重新进行回归。
- estat vif：本命令旨在对新模型进行多重共线性检验。
- regress V2 V3：本命令旨在上步的基础上剔除 P 值不显著的变量后再重新进行回归。

03 设置完毕后，按键盘上的回车键，等待输出结果。

11.3.4　结果分析

在 Stata 14.0 主界面的结果窗口可以看到如图 11.44~图 11.52 所示的分析结果。

1．对数据进行描述性分析的结果

图 11.44 是对数据进行描述性分析的结果。关于这一分析过程对于回归分析的重要意义已在前面章节中论述过，此处不再重复讲解。

```
. summarize  V1 V2 V3 V4 V5 V6,detail

                             V1
-------------------------------------------------------------
      Percentiles      Smallest
 1%         1996           1996
 5%         1996           1997
10%         1996           1998       Obs                   8
25%       1997.5           1999       Sum of Wgt.           8

50%       1999.5                      Mean             1999.5
                        Largest       Std. Dev.       2.44949
75%       2001.5           2000
90%         2003           2001       Variance              6
95%         2003           2002       Skewness              0
99%         2003           2003       Kurtosis       1.761905

                             V2
-------------------------------------------------------------
      Percentiles      Smallest
 1%      67884.6        67884.6
 5%      67884.6        74462.6
10%      67884.6          78345       Obs                   8
25%      76403.8          82067       Sum of Wgt.           8

50%      85754.5                      Mean           88992.51
                        Largest       Std. Dev.      16681.17
75%     101243.5          89442
90%     117251.9          97315       Variance       2.78e+08
95%     117251.9         105172       Skewness       .4398428
99%     117251.9       117251.9       Kurtosis       2.043855

                             V3
-------------------------------------------------------------
      Percentiles      Smallest
 1%        36590          36590
 5%        36590          38089
10%        36590          38385       Obs                   8
25%        38237          40568       Sum of Wgt.           8

50%      42444.5                      Mean              43776
                        Largest       Std. Dev.      6420.092
75%        49198          44321
90%        53859          47710       Variance       4.12e+07
95%        53859          50686       Skewness       .3874834
99%        53859          53859       Kurtosis       1.683573

                             V4
-------------------------------------------------------------
      Percentiles      Smallest
 1%           10             10
 5%           10           10.5
10%           10           11.6       Obs                   8
25%        11.05           12.5       Sum of Wgt.           8

50%         13.1                      Mean              12.85
                        Largest       Std. Dev.      2.174528
75%         13.9           13.7
90%         16.7           13.8       Variance       4.728571
95%         16.7             14       Skewness        .325807
99%         16.7           16.7       Kurtosis       2.349168

                             V5
-------------------------------------------------------------
      Percentiles      Smallest
 1%        10813          10813
 5%        10813          11356
10%        10813          11670       Obs                   8
25%        11513          12393       Sum of Wgt.           8

50%      12974.5                      Mean           13780.25
                        Largest       Std. Dev.      2882.102
75%        15674          13556
90%        19106          14808       Variance        8306510
95%        19106          16540       Skewness       .7700467
99%        19106          19106       Kurtosis       2.364367

                             V6
-------------------------------------------------------------
      Percentiles      Smallest
 1%        15733          15733
 5%        15733          16000
10%        15733          16074       Obs                   8
25%        16037          16100       Sum of Wgt.           8

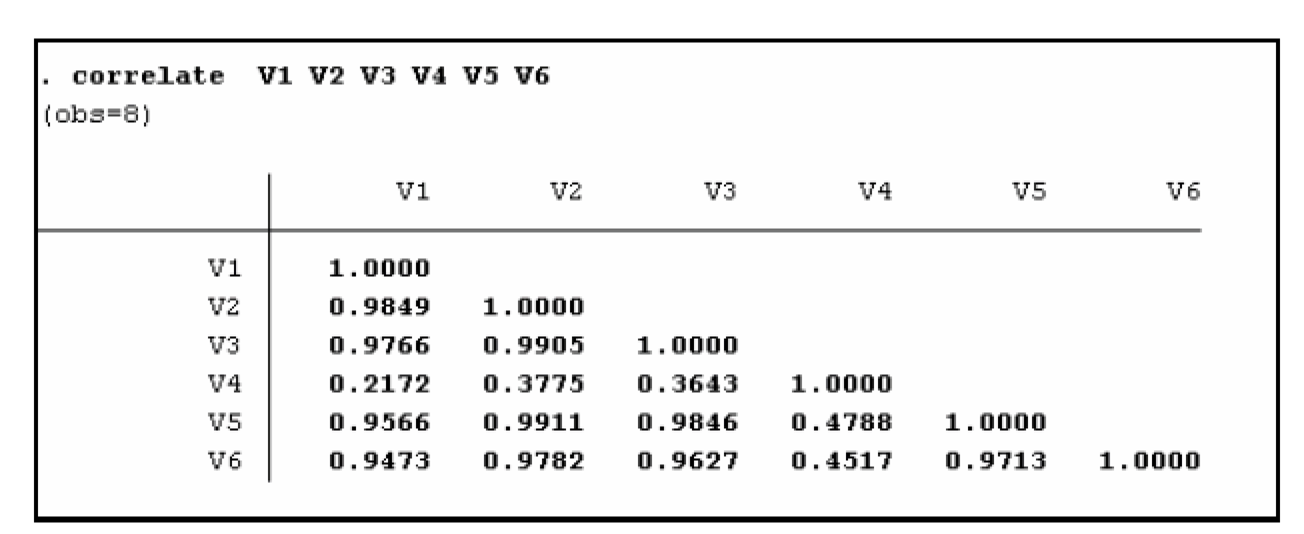
50%        16200                      Mean           16282.88
                        Largest       Std. Dev.      397.3187
75%        16548          16300
90%        16960          16396       Variance       157862.1
95%        16960          16700       Skewness       .4391698
99%        16960          16960       Kurtosis       2.237363
```

图 11.44　对数据进行描述性分析的结果

在如图 11.44 所示的分析结果中可以看出，数据的总体质量还是可以的，没有极端异常值，变量间的量纲差距、变量的偏度、峰度也是可以接受的，可以进行下一步的分析。

2．对数据进行相关性分析的结果

图 11.45 是对数据进行相关性分析的结果。关于这一分析过程对于回归分析的重要意义已在前面章节中论述过，此处不再重复讲解。



. correlate V1 V2 V3 V4 V5 V6
(obs=8)

	V1	V2	V3	V4	V5	V6
V1	1.0000					
V2	0.9849	1.0000				
V3	0.9766	0.9905	1.0000			
V4	0.2172	0.3775	0.3643	1.0000		
V5	0.9566	0.9911	0.9846	0.4788	1.0000	
V6	0.9473	0.9782	0.9627	0.4517	0.9713	1.0000

图 11.45　对数据进行相关性分析的结果

在图 11.45 中，变量间的相关系数非常大，这意味着变量间存在很高程度的信息重叠，模型很有可能存在多重共线性问题。

3．对数据进行回归分析的结果

图 11.46 是对数据进行回归分析的结果。

```
. regress  V2 V3 V4 V5 V6

      Source |       SS       df       MS              Number of obs =       8
-------------+------------------------------           F(  4,     3) =  348.28
       Model |  1.9436e+09     4   485910915           Prob > F      =  0.0002
    Residual |  4185548.75     3  1395182.92           R-squared     =  0.9979
-------------+------------------------------           Adj R-squared =  0.9950
       Total |  1.9478e+09     7   278261315           Root MSE      =  1181.2

------------------------------------------------------------------------------
          V2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          V3 |   .0040429   .5633146     0.01   0.995    -1.788676    1.796761
          V4 |  -931.3118   327.7201    -2.84   0.066    -1974.263    111.6399
          V5 |   4.686809   1.391856     3.37   0.043     .2573033    9.116316
          V6 |   10.28367   4.790103     2.15   0.121    -4.960572    25.52792
       _cons |  -131250.3   68579.04    -1.91   0.152    -349499.4    86998.81
------------------------------------------------------------------------------
```

图 11.46　对数据进行回归分析的结果

从上述分析结果中可以看出共有 8 个样本参与了分析，模型的 F 值(4, 3) = 348.28，P 值（Prob > F） = 0.0002，说明模型整体上是非常显著的。模型的可决系数（R-squared）= 0.9979，模型修正的可决系数（Adj R-squared）= 0.9950，说明模型的解释能力非常不错。

模型的回归方程是：

V2=0.0040429*V3-931.3118*V4+4.686809*V5+10.28367*V6-131250.3

变量 V3 的系数标准误是 0.5633146，t 值为 0.01，P 值为 0.995，系数是非常不显著的，95%的置信区间为[-1.788676, 1.796761]。变量 V4 的系数标准误是 327.7201，t 值为-2.84，P 值为 0.066，系数的显著程度不高，95%的置信区间为[-1974.263, 111.6399]。变量 V5 的系数标准误是 1.391856，t 值为 3.37，P 值为 0.043，系数是非常显著的，95%的置信区间为[0.2573033, 9.116316]。变量 V6 的系数标准误是 4.790103，t 值为 2.15，P 值为 0.121，系数是非常不显著的，95%的置信区间为[-4.960572, 25.52792]。常数项的系数标准误是 68579.04，t 值为-1.91，P 值为 0.152，系数也是非常不显著的，95%的置信区间为[-349499.4, 86998.81]。

从上面的分析可以看出，国内生产总值与货物周转量、原煤、发电量、原油等变量进行回归得到的模型中部分变量的系数非常不显著，而且原煤产量的系数居然是负值，这显然是不符合现实情况的，造成这些现象的根源就在于模型存在着程度比较高的多重共线性问题。

4．对模型进行多重共线性检验的结果

图 11.47 是对模型进行多重共线性检验的结果。

```
. estat vif

    Variable |       VIF       1/VIF
-------------+----------------------
          V5 |     80.74    0.012386
          V3 |     65.62    0.015239
          V6 |     18.17    0.055026
          V4 |      2.55    0.392461
-------------+----------------------
    Mean VIF |     41.77
```

图 11.47　对模型进行多重共线性检验的结果

从图 11.47 中可以看出，Mean VIF 的值是 41.77，远远大于合理值 10，所以模型存在较高程度的多重共线性，其中 V5 的方差膨胀因子最高，即 80.74，所以需要将 V5 剔除以后重新进

行回归。

图 11.48 是在上步的基础上剔除最大的方差膨胀因子再重新进行回归的结果。

```
. regress  V2 V3 V4 V6

      Source |       SS       df       MS              Number of obs =       8
-------------+------------------------------           F(  3,     4) =  128.49
       Model |  1.9278e+09     3  642607998            Prob > F      =  0.0002
    Residual |  20005214.2     4  5001303.55           R-squared     =  0.9897
-------------+------------------------------           Adj R-squared =  0.9820
       Total |  1.9478e+09     7  278261315            Root MSE      =  2236.4

------------------------------------------------------------------------------
          V2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          V3 |   1.671362   .5085665     3.29   0.030     .2593548    3.083369
          V4 |  -182.1422   455.5875    -0.40   0.710    -1447.056    1082.771
          V6 |    15.5194   8.578151     1.81   0.145    -8.297364    39.33617
       _cons |    -234533   116132.3    -2.02   0.114    -556967.8    87901.88
------------------------------------------------------------------------------
```

图 11.48　重新进行回归的结果

关于本结果的详细解读方式，前面多有提及，限于篇幅不再赘述。

图 11.49 是对新模型进行多重共线性检验的结果。

```
. estat vif

    Variable |       VIF       1/VIF
-------------+----------------------
          V6 |     16.26    0.061506
          V3 |     14.92    0.067020
          V4 |      1.37    0.727967
-------------+----------------------
    Mean VIF |     10.85
```

图 11.49　对新模型进行多重共线性检验的结果

从图 11.49 中可以看出，Mean VIF 的值是 10.85，接近合理值 10，所以模型的多重共线性得到了很大程度的改善，下面剔除目前最大的方差膨胀因子 V6，继续进行回归。

图 11.50 是在上步的基础上剔除最大的方差膨胀因子再重新进行回归的结果。

```
. regress  V2 V3 V4

      Source |       SS       df       MS              Number of obs =       8
-------------+------------------------------           F(  2,     5) =  131.37
       Model |  1.9115e+09     2  955727052            Prob > F      =  0.0000
    Residual |  36375104.5     5  7275020.9            R-squared     =  0.9813
-------------+------------------------------           Adj R-squared =  0.9739
       Total |  1.9478e+09     7  278261315            Root MSE      =  2697.2

------------------------------------------------------------------------------
          V2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          V3 |   2.555185   .1705049    14.99   0.000     2.116889    2.993482
          V4 |   148.2452   503.3999     0.29   0.780    -1145.785    1442.276
       _cons |  -24768.24    7955.57    -3.11   0.026    -45218.68   -4317.793
------------------------------------------------------------------------------
```

图 11.50　重新进行回归的结果

关于本结果的详细解读方式，前面多有提及，限于篇幅不再赘述。

图 11.51 是对新模型进行多重共线性检验的结果。

从图 11.51 中可以看出，Mean VIF 的值是 1.15，远远小于合理值 10，所以模型的多重共线性得到了很大程度的改善。但是根据图 11.50 所示的结果，V4 的系数并不显著，可以把 V4 也剔除，再重新进行回归。

```
. estat vif

    Variable |       VIF       1/VIF
-------------+----------------------
          V3 |      1.15    0.867321
          V4 |      1.15    0.867321
-------------+----------------------
    Mean VIF |      1.15
```

图 11.51　对新模型进行多重共线性检验的结果

图 11.52 是在上步的基础上剔除系数不显著的变量再重新进行回归的结果。

```
. regress  V2 V3

      Source |       SS           df       MS      Number of obs   =         8
-------------+----------------------------------   F(  1,      6)  =    309.81
       Model |  1.9108e+09         1  1.9108e+09   Prob > F        =    0.0000
    Residual |  37006017.3         6  6167669.55   R-squared       =    0.9810
-------------+----------------------------------   Adj R-squared   =    0.9778
       Total |  1.9478e+09         7   278261315   Root MSE        =    2483.5

------------------------------------------------------------------------------
          V2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          V3 |   2.573475   .1462077    17.60   0.000     2.215718    2.931232
       _cons |  -23663.93   6460.335    -3.66   0.011    -39471.81   -7856.063
------------------------------------------------------------------------------
```

图 11.52　重新进行回归的结果

从图 11.52 中可以看出，模拟的整体显著性、模型的解释能力、模型中各变量和常数项的系数显著性都达到了近乎完美的状态。最终的结论是参与分析的变量中，货物周转量能够最大程度地解释国内生产总值，货物周转量越大，国内生产总值也就越大。

11.3.5　案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

下面使用因子分析方法解决模型的多重共线性问题。

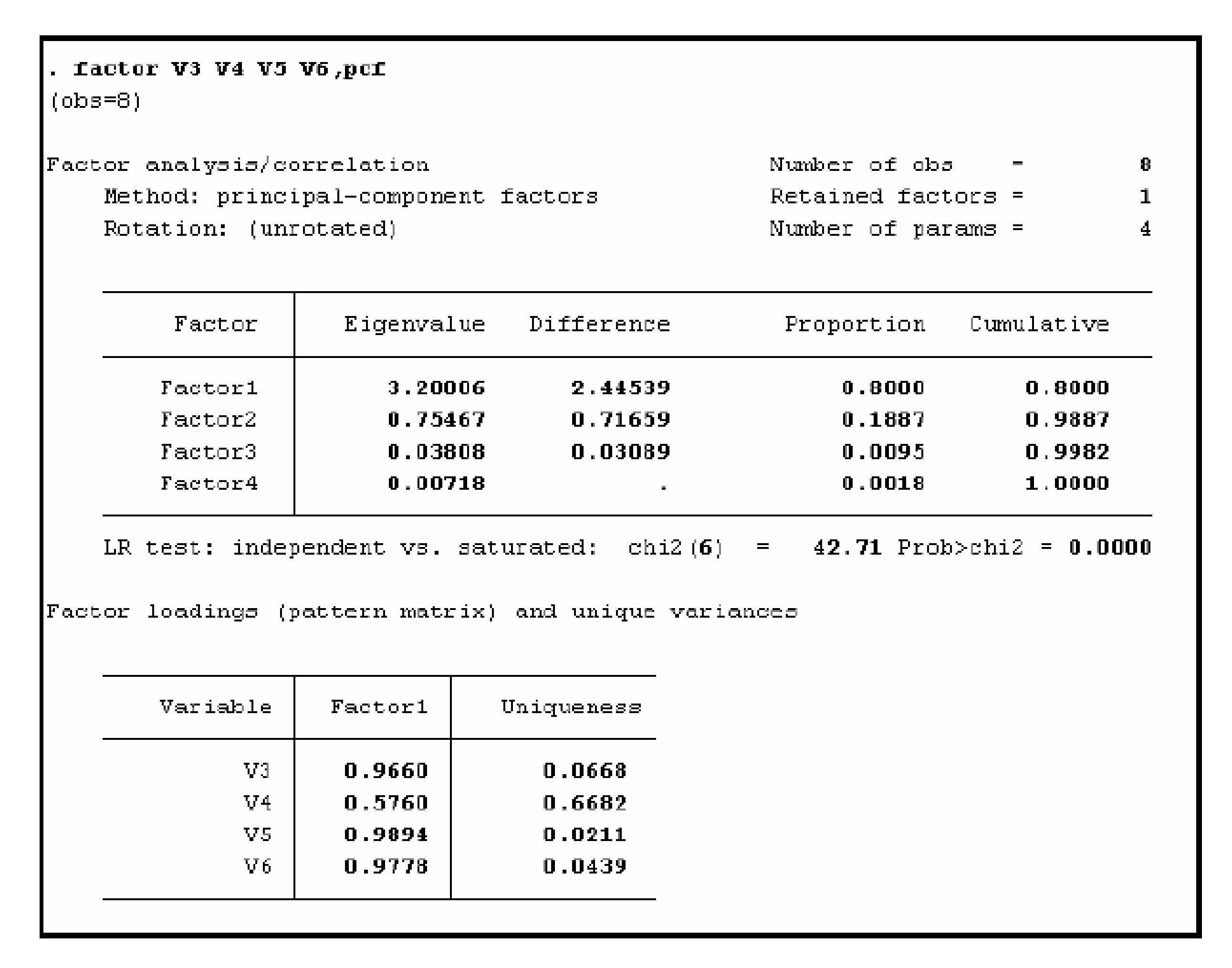
以本例为例进行说明，操作命令如下。

- factor V3 V4 V5 V6,pcf：本命令旨在对 V3、V4、V5、V6 变量提取公因子。
- predict f1：本命令旨在产生已提取的公因子变量 f1。
- reg V2 f1：本命令旨在以 V2 为因变量，以 f1 为自变量进行最小二乘回归分析。
- vif：本命令旨在对模型进行多重共线性检验。

在命令窗口输入命令并按回车键进行确认，结果如图 11.53~图 11.56 所示。

图 11.53 是对 V3、V4、V5、V6 变量提取公因子的结果。对本结果的解读已有详细表述，此处限于篇幅不再赘述。

图 11.54 是因子分析得到的公因子变量 f1 以及因子得分系数情况。



```
. factor V3 V4 V5 V6,pcf
(obs=8)

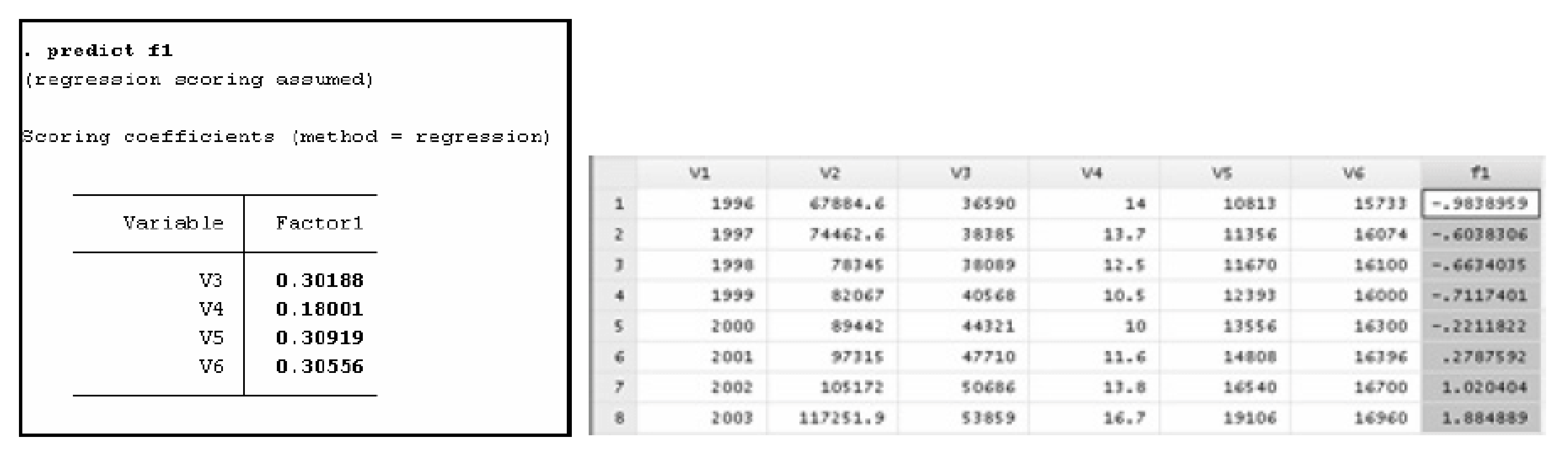
Factor analysis/correlation                    Number of obs    =        8
    Method: principal-component factors        Retained factors =        1
    Rotation: (unrotated)                      Number of params =        4

    --------------------------------------------------------------------------
         Factor  |   Eigenvalue   Difference        Proportion   Cumulative
    -------------+------------------------------------------------------------
        Factor1  |      3.20006      2.44539            0.8000       0.8000
        Factor2  |      0.75467      0.71659            0.1887       0.9887
        Factor3  |      0.03808      0.03089            0.0095       0.9982
        Factor4  |      0.00718            .            0.0018       1.0000
    --------------------------------------------------------------------------
    LR test: independent vs. saturated:  chi2(6)  =   42.71 Prob>chi2 = 0.0000

Factor loadings (pattern matrix) and unique variances

    -------------------------------------
        Variable |  Factor1 | Uniqueness
    -------------+----------+------------
              V3 |   0.9660 |     0.0668
              V4 |   0.5760 |     0.6682
              V5 |   0.9894 |     0.0211
              V6 |   0.9778 |     0.0439
    -------------------------------------
```

图 11.53　对 V3、V4、V5、V6 变量提取公因子的结果



```
. predict f1
(regression scoring assumed)

Scoring coefficients (method = regression)

    ------------------------
        Variable |  Factor1
    -------------+----------
              V3 |  0.30188
              V4 |  0.18001
              V5 |  0.30919
              V6 |  0.30556
    ------------------------
```

	V1	V2	V3	V4	V5	V6	f1
1	1996	67884.6	36590	14	10813	15733	-.9838959
2	1997	74462.6	38385	13.7	11356	16074	-.6038306
3	1998	78345	38089	12.5	11670	16100	-.6634035
4	1999	82067	40568	10.5	12393	16000	-.7117401
5	2000	89442	44321	10	13556	16300	-.2211822
6	2001	97315	47710	11.6	14808	16396	.2787592
7	2002	105172	50686	13.8	16540	16700	1.020404
8	2003	117251.9	53859	16.7	19106	16960	1.884889

图 11.54　因子得分系数矩阵

根据图 11.54 展示的因子得分系数矩阵，可以写出公因子的表达式。值得一提的是，在表达式中各个变量已经不是原始变量，而是标准化变量。

表达式如下：

f1= 0.30188*货物周转量+ 0.18001*原煤+ 0.30919*发电量+ 0.30556*原油

图 11.55 是以 V2 为因变量、以 f1 为自变量进行最小二乘回归分析的结果。

```
. reg V2 f1

      Source |       SS       df       MS              Number of obs =       8
-------------+------------------------------           F(  1,     6) =  103.71
       Model |  1.8413e+09     1  1.8413e+09           Prob > F      =  0.0001
    Residual |   106524045     6  17754007.5           R-squared     =  0.9453
-------------+------------------------------           Adj R-squared =  0.9362
       Total |  1.9478e+09     7   278261315           Root MSE      =  4213.6

------------------------------------------------------------------------------
          V2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          f1 |   16218.62   1592.572    10.18   0.000     12321.73     20115.5
       _cons |   88992.51   1489.715    59.74   0.000     85347.31    92637.71
------------------------------------------------------------------------------
```

图 11.55　以 V2 为因变量、以 f1 为自变量进行最小二乘回归分析的结果

从图 11.55 中可以看出，模拟的整体显著性、模型的解释能力、模型中各变量和常数项的系数显著性都达到了近乎完美的状态。

图 11.56 是对模型进行多重共线性检验的结果。

```
. vif

    Variable |       VIF       1/VIF
-------------+----------------------
          f1 |      1.00    1.000000
-------------+----------------------
    Mean VIF |      1.00
```

图 11.56 对新模型进行多重共线性检验的结果

从图 11.56 中可以看出，Mean VIF 的值是 1，远远小于合理值 10，所以模型的多重共线性得到了很大程度的改善。

11.4 本章习题

（1）某著名跨国公司拥有自己的一套职员评价体系，搜集并整理了公司内部 133 名职员的相关数据，如表 11.4 所示。表中的内容包括职员的年薪、工作年限、学历职称、工作能力、敬业精神 5 部分的内容，试使用职员年薪作为因变量，以职员的工作年限、学历职称、工作能力、敬业精神作为自变量，对这些数据使用最小二乘回归分析的方法进行研究，并进行异方差检验，最终建立合适的回归方程模型用于描述变量之间的关系。

表 11.4 某著名跨国公司搜集整理的 133 名职员的相关数据

编号	职员年薪	工作年限	学历职称	工作能力	敬业精神
1	6.855 409	2.397 895	5.288 267	5.872 118	5.327 876
2	6.514 713	2.564 949	5.323 01	5.860 786	5.010 635
3	6.263 398	2.564 949	5.389 072	5.673 323	5.043 425
4	6.216 606	3.091 043	5.147 495	5.010 635	5.236 442
5	7.085 064	3.218 876	5.342 334	5.187 386	5.135 798
6	6.507 278	3.218 876	5.123 964	5.983 936	5.117 994
…	…	…	…	…	…
130	10.414 93	8.972 844	5.081 404	5.181 784	5.181 784
131	11.075 07	9.038 246	5.446 737	5.765 191	5.293 305
132	10.627 12	9.064 389	5.411 646	5.579 73	5.204 007
133	10.778 81	9.081 029	5.442 418	5.814 131	5.247 024

（2）表 11.5 给出了某旅游景点游客量和资金投入的有关数据，试使用游客量作为因变量，以资金投入作为自变量，对这些数据使用最小二乘回归分析的方法进行研究，并进行自相关检验，最终建立合适的回归方程模型用于描述变量之间的关系。

表 11.5　某旅游景点游客量和资金投入的有关数据

月份	游客量/万人	资金投入/万元
1	21.45	282.9
2	23.01	285.9
3	24.08	290.9
4	25.07	302.9
5	26.99	315.98
6	26.01	310.25
…	…	…
45	65.99	455.99
46	64.01	470.29
47	58.96	473.01
48	57.98	511.99
49	68.99	551

（3）表 11.6 给出了我国 1992－2000 年国民经济主要指标统计数据。试使用国内生产总值作为因变量，以货物周转量、原煤、发电量、原油等作为自变量，对这些数据使用最小二乘回归分析的方法进行研究，并进行多重共线性检验，最终建立合适的回归方程模型用于描述变量之间的关系。

表 11.6　我国 1992－2000 年国民经济主要指标统计数据

年份	国内生产总值/亿元	货物周转量/亿吨千米	原煤/亿吨	发电量/亿千瓦时	原油/万吨
1992	26 638.1	29 218.0	11.2	7 539.0	14 210.0
1993	34 634.4	30 510.0	11.5	8 394.0	14 524.0
1994	46 759.4	33 261.0	12.4	9 281.0	14 608.0
1995	58 478.1	35 730.0	13.6	10 077.0	15 005.0
1996	67 884.6	36 454.0	14.0	10 813.0	15 733.0
1997	74 462.6	38 368.0	13.7	11 356.0	16 074.0
1998	78 345.0	38 046.0	12.5	11 670.0	16 100.0
1999	82 067.0	40 496.0	10.5	12 393.0	16 000.0
2000	89 403.5	44 452.0	10.0	13 556.0	16 300.0

第 12 章　Stata 非线性回归分析

前面讲述的回归分析方法都属于线性回归的范畴，即因变量和自变量之间存在线性关系。在很多情况下，线性模型是对真实情况的一种合理但又简单的近似。如果遇到回归参数不是线性的，也不能通过转换的方法将其转换为线性的参数，又该如何处理呢？这时候就需要用到本章将要讲述的非线性回归分析。常用的非线性分析方法有 3 种，包括非参数回归分析、转换变量回归分析以及非线性回归分析。下面就以实例的方式一一介绍这几种方法在 Stata 中的应用。

12.1　实例——非参数回归分析

12.1.1　非参数回归分析的功能与意义

非参数回归分析（Nonparametric Methods）与前面讲述的回归方式区别很大，是一种探索性工具，通常不会像其他回归方法一样形成一个明确的回归方程，基本上是展示因变量与自变量之间关系的图形工具。其优势在于在不要求研究者事先设定模型的情况下就可直观、概要地描述数据。

12.1.2　相关数据来源

	下载资源:\video\chap12\…
	下载资源:\sample\chap12\案例12.1.dta

【例 12.1】某国内保险公司采取区域事业部制的组织机构模式，在国内有两个事业部：北方事业部和南方事业部。该公司对其客户经理制定了严格的激励约束措施，客户经理的薪酬为基本工资乘以绩效考核系数，绩效考核系数上不封顶、下不保底，所以客户经理之间的收入差距很大。某研究者随机抽取的部分客户经理的历年考核系数如表 12.1 所示，请用非参数回归方法研究年份和绩效考核系数两个变量之间的关系。

表 12.1　某国内保险公司客户经理绩效考核系数表

所属事业部	年份	绩效考核系数
北方事业部	2000	1.8
北方事业部	2000	2
北方事业部	2000	1.9
北方事业部	2001	1.7
北方事业部	2001	1.6

（续表）

所属事业部	年份	绩效考核系数
…	…	…
南方事业部	2010	1.49
南方事业部	2010	1.69
南方事业部	2010	1.92

12.1.3 Stata 分析过程

在用 Stata 进行分析之前，要把数据录入到 Stata 中。本例中有 3 个变量，分别为所属事业部、年份和绩效考核系数。把所属事业部变量设定为 region，并且把北方事业部设定为 1，把南方事业部设定为 2，把年份变量定义为 year，把绩效考核系数定义为 coefficient，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作在第 1 章中已有详细讲述。录入完成后数据如图 12.1 所示。

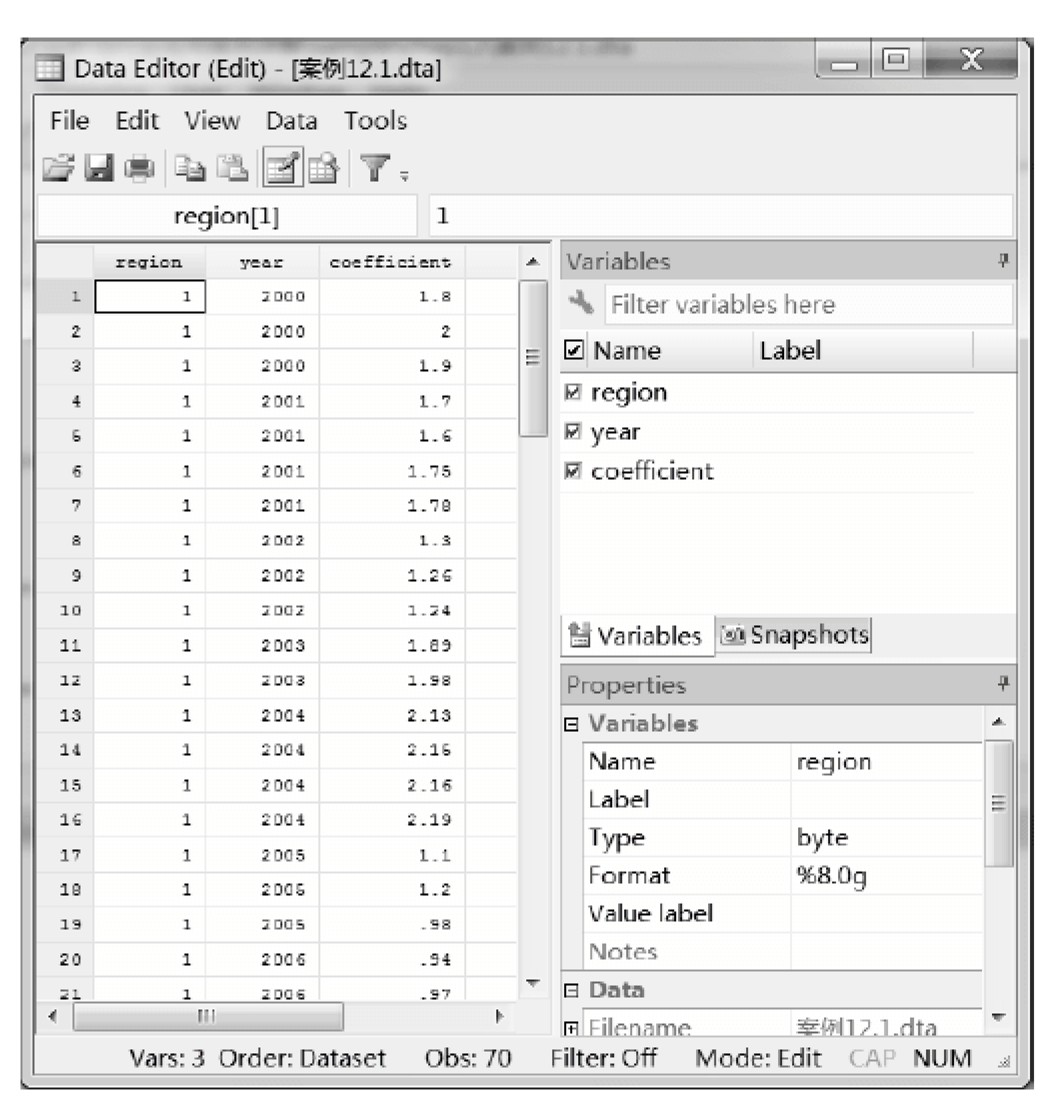


图 12.1 案例 12.1 数据

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入如下命令。

- summarize year coefficient,detail：本命令的含义是对年份和绩效考核系数进行描述性分析，简要探索数据特征，从整体上对数据有一个清晰直观的把握。
- twoway line coefficient year：本命令的含义是对运用 Stata 的制图功能，描述年份和绩效考核系数之间的变化关系。
- graph twoway mband coefficient year || scatter coefficient year: 本命令的含义是对数据进行非参数回归并且绘制年份和绩效考核系数之间的散点图。
- graph twoway mband coefficient year || scatter coefficient year || ,by(region)：本命令的含

义是以事业部为分类对数据进行非参数回归，并且绘制年份和绩效考核系数之间的散点图。

- lowess　coefficient year if region==1：本命令是对数据进行修匀，这是非参数回归的另外一种重要形式。
- graph twoway lowess coefficient year if region==1 || scatter coefficient year：本命令旨在把修匀命令融合到非参数回归中。

03 设置完毕后，按键盘上的回车键，等待输出结果。

12.1.4　结果分析

在 Stata 14.0 主界面的结果窗口可以看到如图 12.2~图 12.7 所示的分析结果。

1．对数据进行描述性分析的结果

图 12.2 是对数据进行描述性分析的结果。关于这一分析过程对于回归分析的重要意义在前面章节中已经论述过，此处不再重复讲解。

```
. summarize year coefficient,detail

                            year
-------------------------------------------------------------
      Percentiles      Smallest
 1%         2000           2000
 5%         2000           2000
10%         2001           2000       Obs                  70
25%         2002           2000       Sum of Wgt.          70

50%         2005                      Mean           2004.971
                        Largest       Std. Dev.        3.1713
75%         2008           2010
90%         2009           2010       Variance       10.05714
95%         2010           2010       Skewness      -.0176288
99%         2010           2010       Kurtosis       1.701294

                         coefficient
-------------------------------------------------------------
      Percentiles      Smallest
 1%          .73            .73
 5%          .89            .75
10%          .98            .84       Obs                  70
25%         1.24            .89       Sum of Wgt.          70

50%         1.78                      Mean           1.735429
                        Largest       Std. Dev.      .5549636
75%          2.1            2.7
90%        2.495            2.8       Variance       .3079846
95%          2.7           2.86       Skewness       .0567911
99%          2.9            2.9       Kurtosis        2.24893
```

图 12.2　对数据进行描述性分析

在如图 12.2 所示的分析结果中，可以得到很多信息，包括百分位数、4 个最小值、4 个最大值、平均值、标准差、偏度、峰度等。

（1）百分位数（Percentiles）

可以看出变量 year 的第 1 个四分位数（25%）是 2002，第 2 个四分位数（50%）是 2005，第 3 个四分位数（75%）是 2008；变量 coefficient 的第 1 个四分位数（25%）是 1.24，第 2 个四分位数（50%）是 1.78，第 3 个四分位数（75%）是 2.1。

（2）4 个最小值（Smallest）

变量 year 最小的 4 个数据值分别是 2000、2000、2000、2000。

变量 coefficient 最小的 4 个数据值分别是 0.73、0.75、0.84、0.89。

（3）4 个最大值（Largest）

变量 year 最大的 4 个数据值分别是 2010、2010、2010、2010。

变量 coefficient 最大的 4 个数据值分别是 2.7、2.8、2.86、2.9。

（4）平均值（Mean）和标准差（Std. Dev）

变量 year 的平均值为 2004.971，标准差是 3.1713。

变量 coefficient 的平均值为 1.735429，标准差是 0.5549636。

（5）偏度（Skewness）和峰度（Kurtosis）

变量 year 的偏度为-0.0176288，为负偏度但不大。

变量 coefficient 的偏度为 0.0567911，为正偏度但不大。

变量 year 的峰度为 1.781294，有一个比正态分布略短的尾巴。

变量 coefficient 的峰度为 2.24893，有一个比正态分布略短的尾巴。

综上所述，数据的总体质量还是可以的，没有极端异常值，变量间的量纲差距、变量的偏度、峰度也是可以接受的，可以进行下一步的分析。

2．描述年份和绩效考核系数之间的关系图

图 12.3 是运用 Stata 的制图功能描述年份和绩效考核系数之间变化关系的结果。

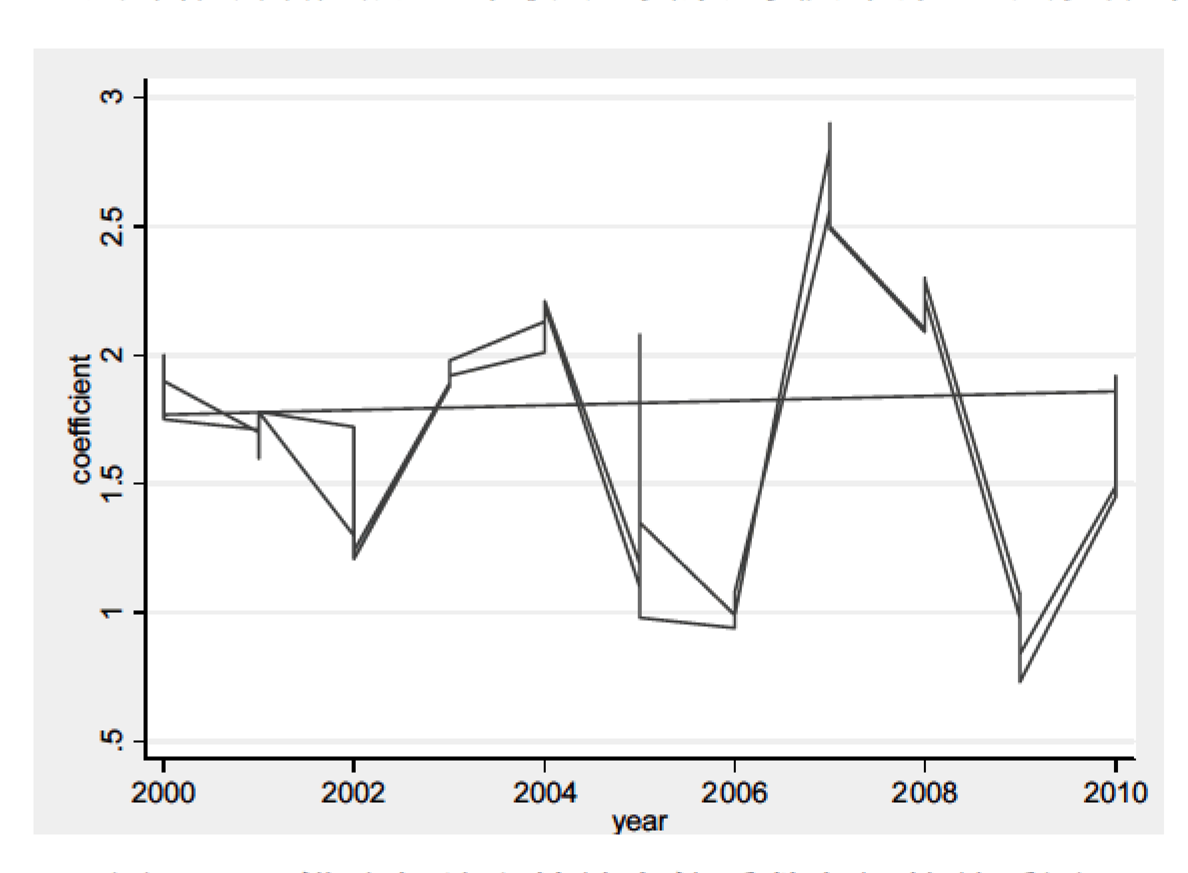


图 12.3　描述年份和绩效考核系数之间的关系图

从图 12.3 中可以看出使用普通的绘图方式来描述年份和绩效考核系数之间的变化关系是非常不清晰的，所以很有必要进行非参数回归来描述这种关系。

3．绘制散点图

图 12.4 是对数据进行非参数回归并且绘制年份和绩效考核系数之间的散点图的结果。

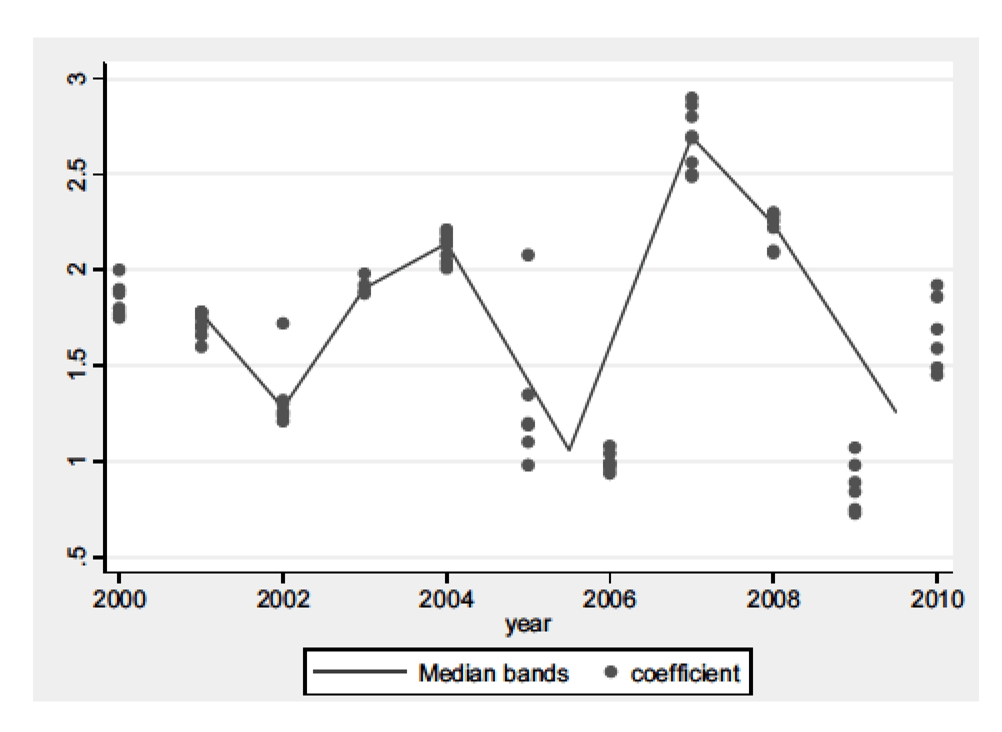


图 12.4　散点图

从图 12.4 可以看出散点图被分成了 8 个垂直等宽的波段，并使用线段将每一波段内的中位数（年份的中位数、绩效考核系数的中位数）连接起来，这条线段直观描绘了绩效考核系数随年份的变化走势。可以认为，绩效考核系数跟年份之间是一种高度波动关系，从 2000 年开始到 2010 年，被观测的客户经理的绩效考核系数先下降又上升，再下降又上升，又下降。

图 12.5 是以事业部为分类，对数据进行非参数回归并且绘制年份和绩效考核系数之间的散点图的结果。

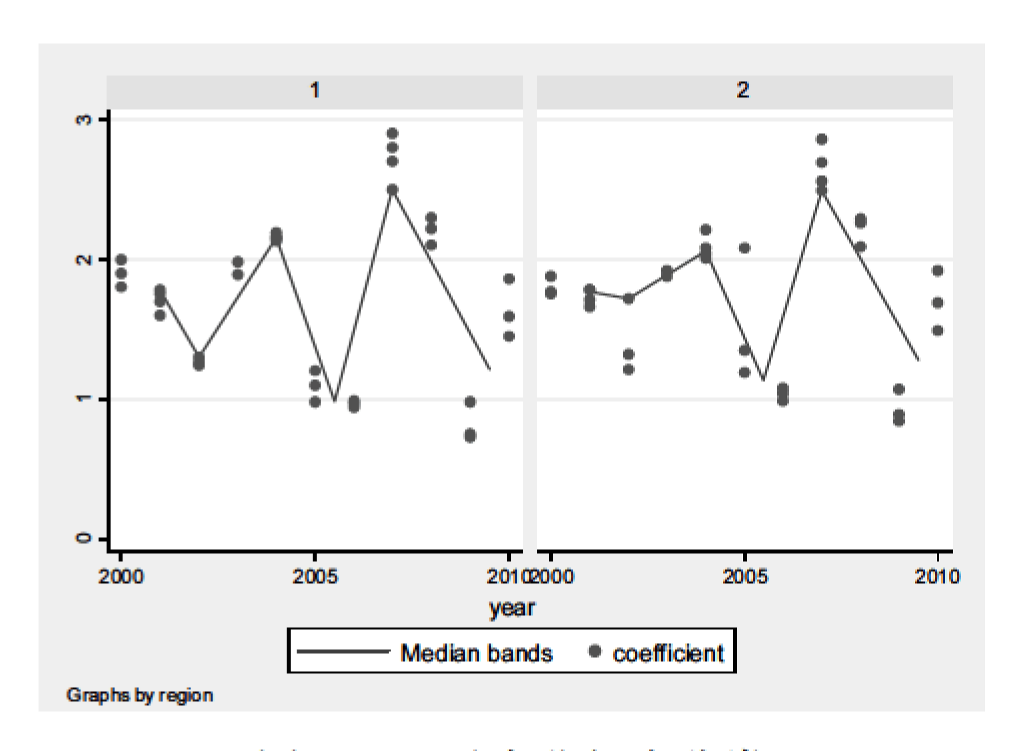


图 12.5　以事业部为分类

从图 12.5 可以看出北方事业部和南方事业部的绩效考核系数的整体走势是很相近的，但是南方事业部的波动要相对平滑一下。

图 12.6 是对数据进行修匀的结果。

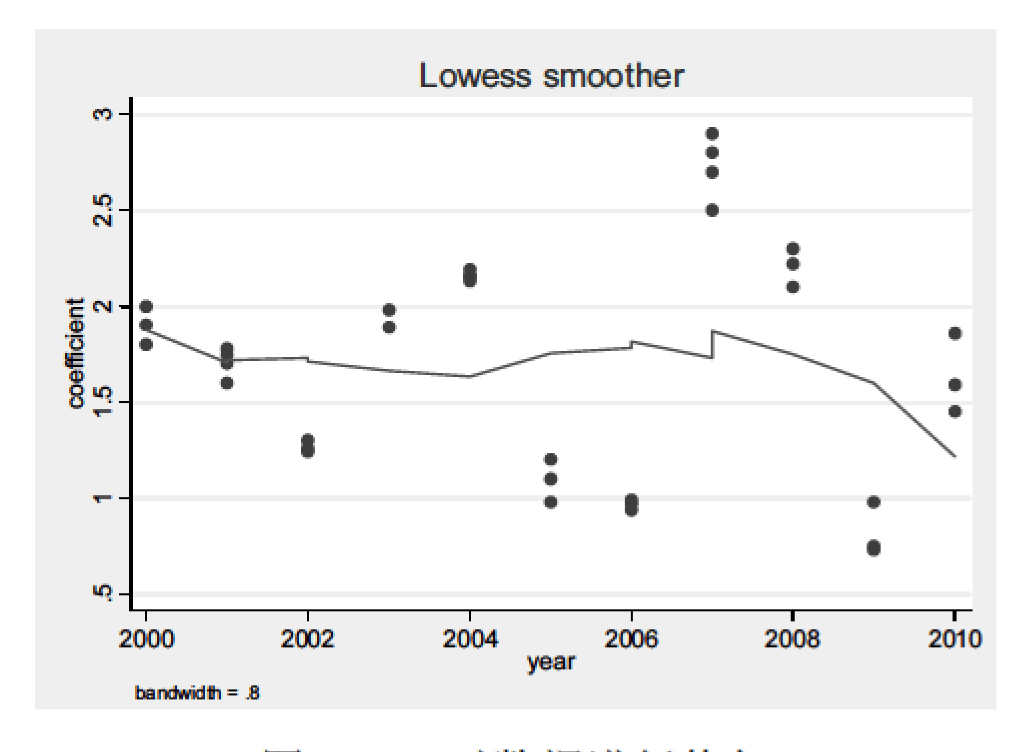


图 12.6　对数据进行修匀

从图 12.6 可以看出，在修匀的情况下绩效考核系数围绕着一条值约为 1.6 的中轴线上下波动。可以初步判定该公司的客户经理的绩效水平是比较高的。

图 12.7 是把修匀命令融合到非参数回归中的结果。

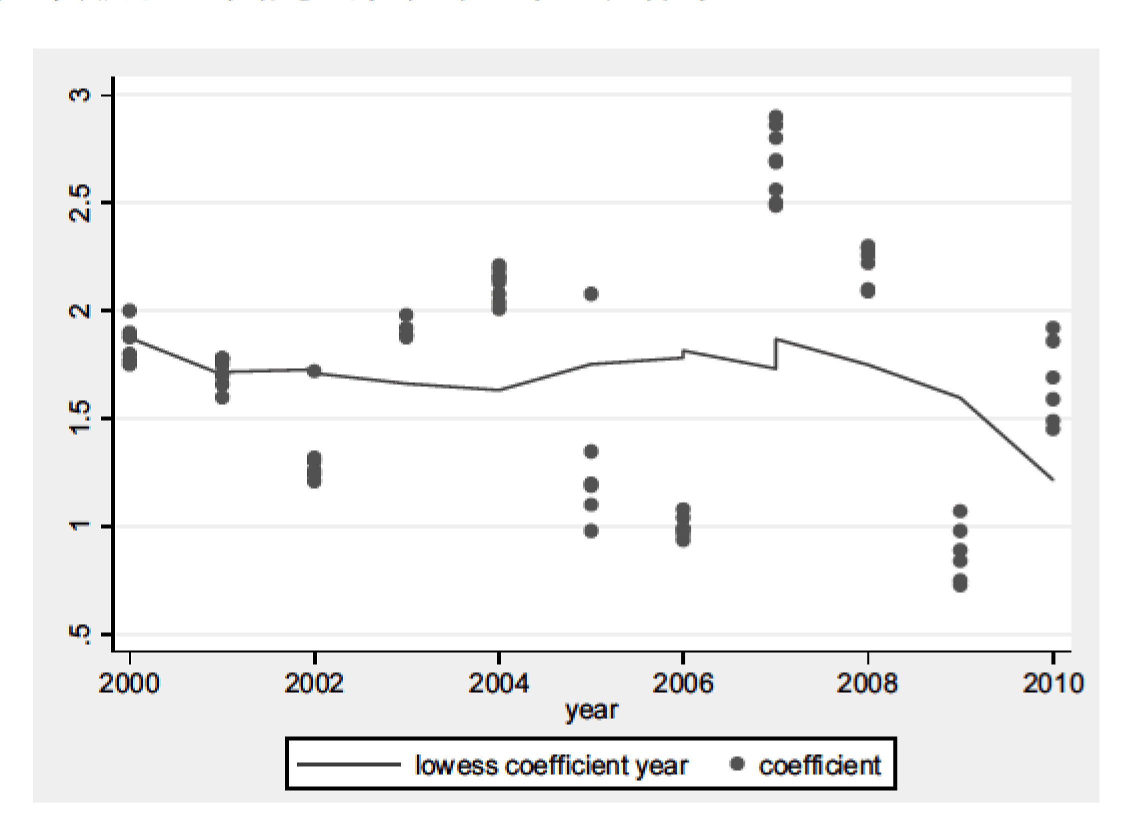


图 12.7　分析结果图

该结果与图 12.6 所示的结果是一致的。

12.1.5　案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

1．延伸 1：设定散点图被分成垂直等宽波段的数量

例如，我们要把散点图分成 10 段垂直等宽的波段，那么操作命令就是：

```
graph twoway mband coefficient year,bands(10) || scatter coefficient year
```

在命令窗口输入命令并按回车键进行确认，结果如图 12.8 所示。

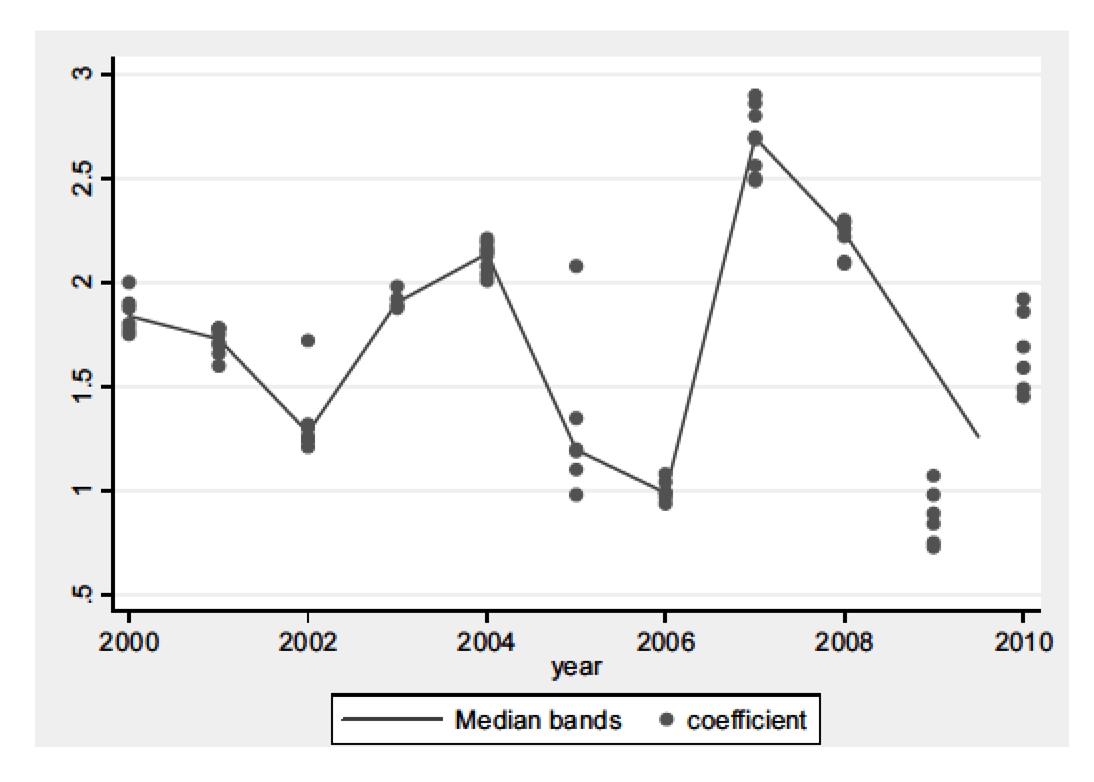


图 12.8　分析结果图

从上面的分析结果中可以看出，相对于系统默认设定，散点图得到了更加细致的划分，绩效考核系数走势也更加清晰明朗。

以事业部为分类对数据进行非参数回归，并且把散点图分成 10 段垂直等宽的波段的操作命令如下：

```
graph twoway mband coefficient year,bands(10) || scatter coefficient
year || ,by(region)
```

在命令窗口输入命令并按回车键进行确认，结果如图 12.9 所示。

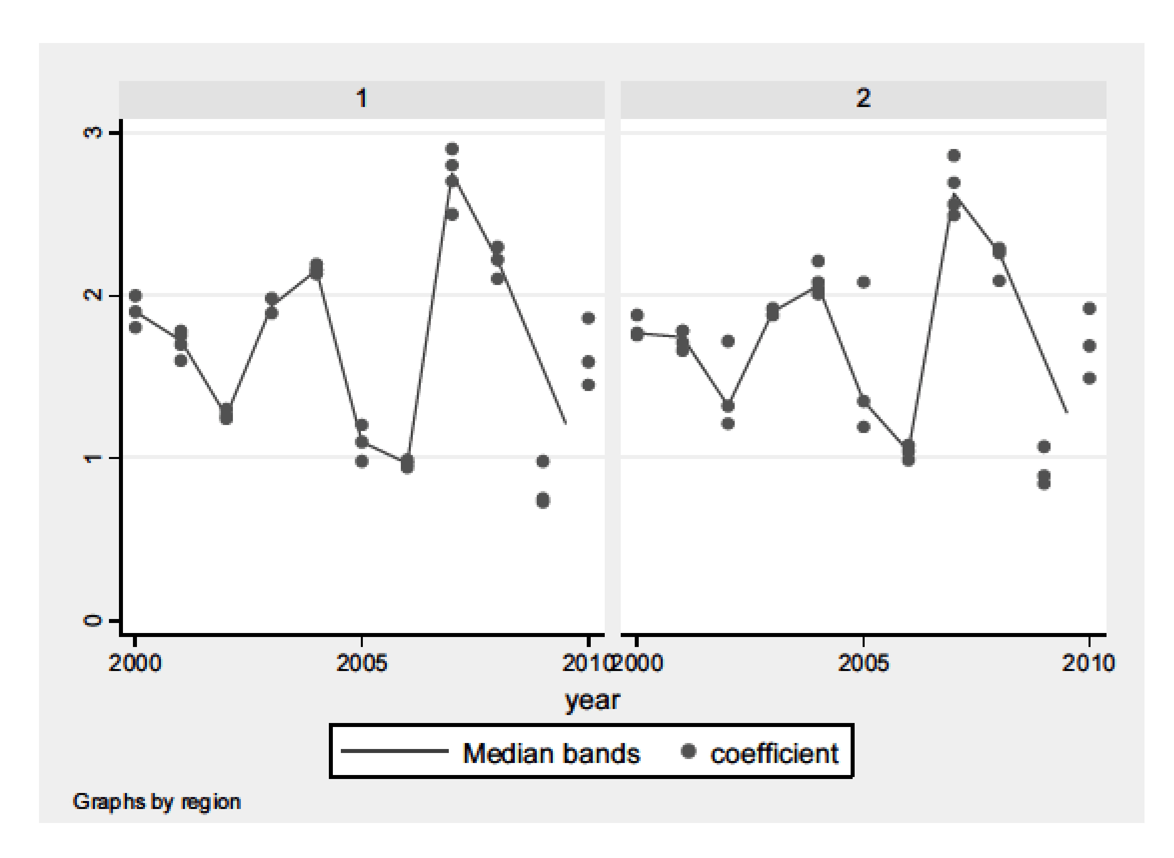


图 12.9　分析结果图

从上面的分析结果中可以看出，相对于系统默认分成的 8 段，散点图得到了更加细致的划分，绩效考核系数走势也更加清晰明朗。

2．延伸 2：设定修匀的波段宽度

例如，要设定对每一点进行修匀的样本比例为 0.4，那么操作命令就是：

```
lowess  coefficient year if region==1,bwidth(0.4)
```

在命令窗口输入命令并按回车键进行确认，结果如图 12.10 所示。

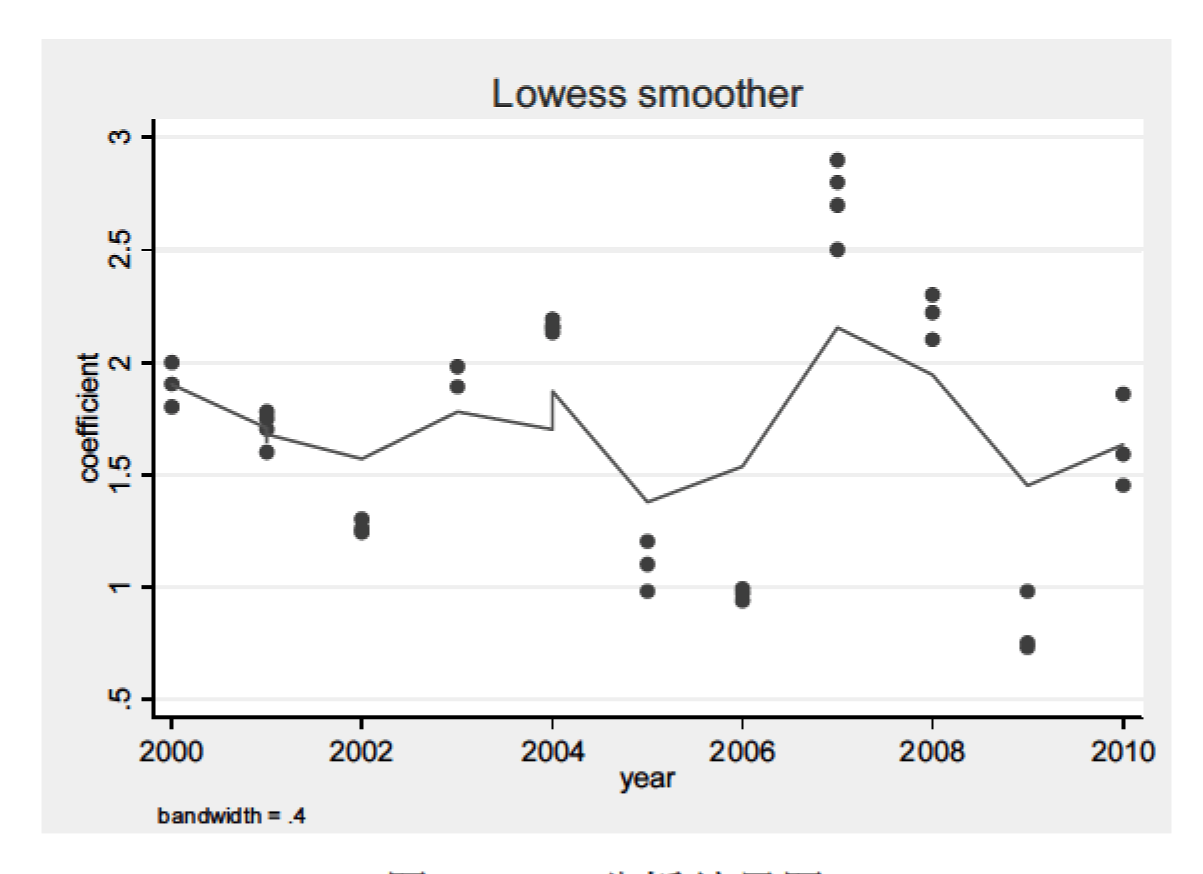


图 12.10　分析结果图

从上面的分析结果中可以看出，数据的波动性得到了增强，修匀程度得到了进一步的降低。

如果设定对每一点进行修匀的样本比例为 0.1，那么操作命令就是：

```
lowess  coefficient year if region==1,bwidth(0.1)
```

在命令窗口输入命令并按回车键进行确认，结果如图 12.11 所示。

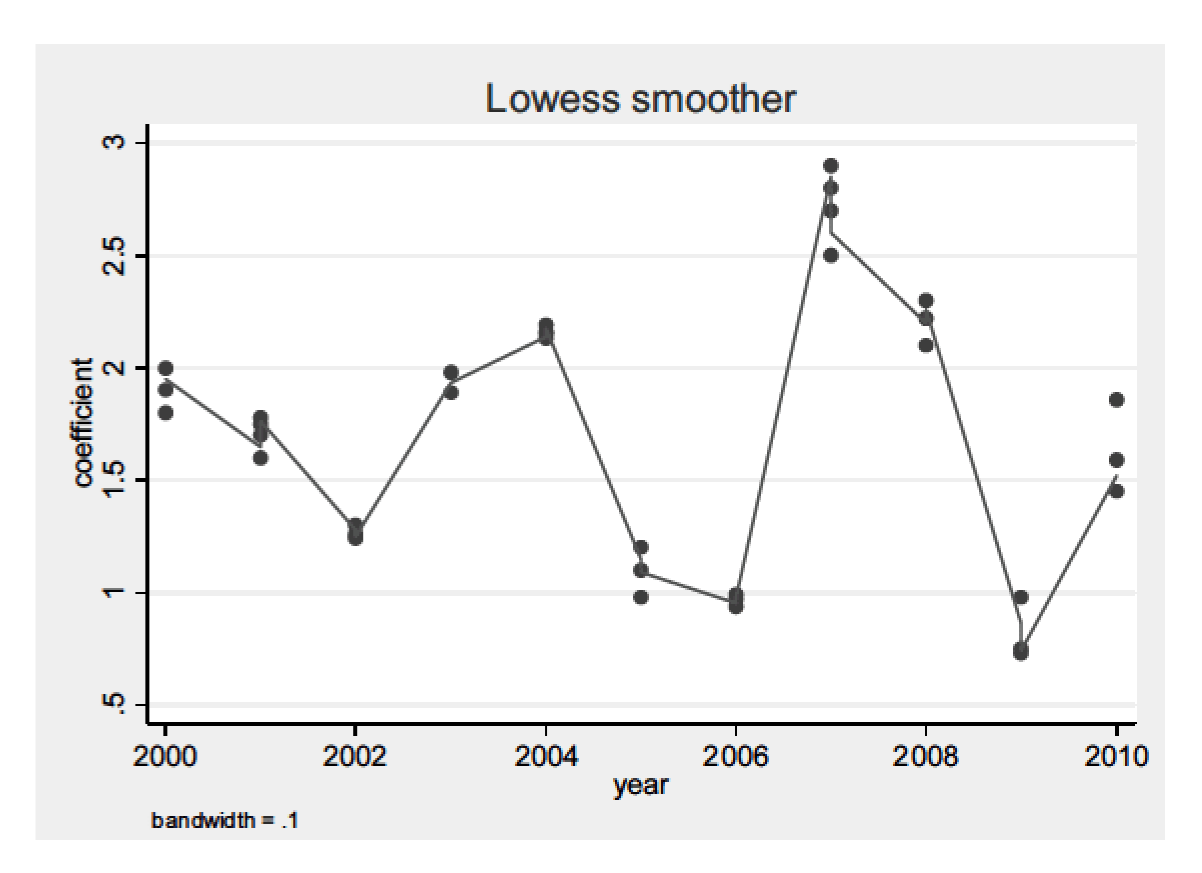


图 12.11　分析结果图

从上面的分析结果可以看出，数据的波动性进一步得到了增强，修匀程度得到了进一步的降低。系统默认的修匀样本比例是 0.8，波段宽度也就是修匀样本比例越接近于 1，数据修匀的程度就越低。

12.2　实例二——转换变量回归分析

12.2.1　转换变量回归分析的功能与意义

转换变量回归分析是解决变量间非线性关系的重要方法之一，基本思想是对一个或者更多的变量进行恰当形式的非线性转换，然后将转换好的变量纳入到线性回归分析模型中进行分析。由此可以看出转换变量回归分析在本质上仍属于线性回归分析的范畴，但它的确是解决描述变量间非线性关系的较好方法。

12.2.2　相关数据来源



	下载资源:\video\chap12\···
	下载资源:\sample\chap12\案例12.2.dta

【例 12.2】研究发现，锡克氏试验阴性率随着儿童年龄的增长而有所升高。山东省某地 1~7 岁儿童锡克氏试验阴性率的资料如表 12.2 所示，试用转换变量回归分析方法拟合曲线。

表 12.2　儿童锡克氏试验阴性率

年龄/岁	阴性率/%
1	56.7
2	75.9
3	90.8
4	93.2
5	96.6
6	95.7
7	96.3

12.2.3　Stata 分析过程

在用 Stata 进行分析之前，要把数据录入到 Stata 中。本例中有两个变量，分别是年龄和阴性率。把年龄变量设定为 age，把阴性率变量设定为 ratio，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作在第 1 章中已有详细讲述。录入完成后数据如图 12.12 所示。

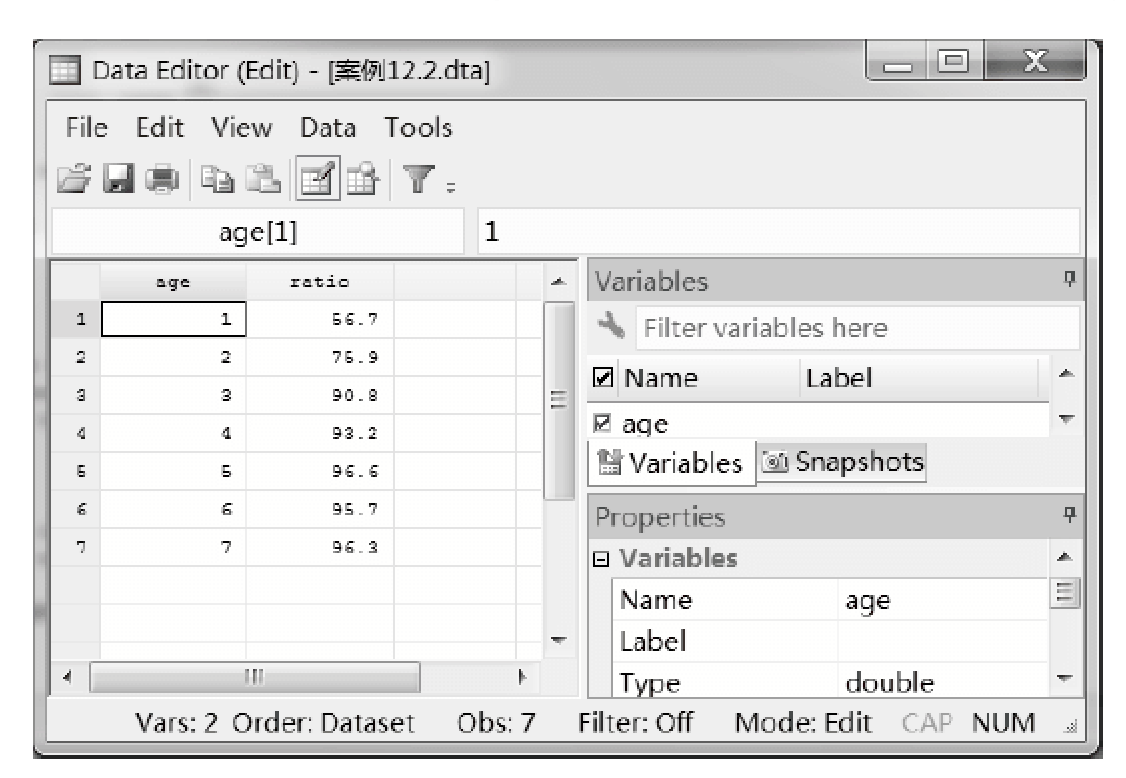


图 12.12　案例 12.2 数据

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入如下操作命令，并按键盘上的回车键进行确认。

- summarize age ratio,detail：本命令的含义是对年龄和阴性率进行描述性分析，简要探索数据特征，从整体上对数据有一个清晰直观的把握。
- twoway line ratio age：本命令旨在通过绘制年龄和阴性率的线形图，从整体上对数据有一个清晰直观的把握。
- graph twoway scatter ratio age || lfit　ratio age：本命令旨在通过绘制年龄和阴性率的散点图，从整体上对数据有一个清晰直观的把握。
- reg ratio age：本命令旨在构建线性模型，以阴性率为因变量，以年龄为自变量，进行最小二乘回归分析，探索变量间的回归关系。
- gen lnage=log(age)：本命令旨在对自变量年龄进行自然对数变换，为下一步的分析做

好准备。

- reg ratio lnage：本命令旨在构建对数模型，以阴性率为因变量，以年龄的对数值为自变量，进行最小二乘回归分析，探索变量间的回归关系。
- gen age2=age^2：本命令旨在对自变量年龄进行二次变换，为下一步的分析做好准备。
- reg ratio age2 age：本命令旨在构建二次模型，以阴性率为因变量，以年龄以及年龄的二次方为自变量，进行最小二乘回归分析，探索变量间的回归关系。
- gen age3=age^3：本命令旨在对自变量年龄进行三次变换，为下一步的分析做好准备。
- reg ratio age3 age2 age：本命令旨在构建三次模型，以阴性率为因变量，以年龄、年龄的二次方以及年龄的三次方为自变量，进行最小二乘回归分析，探索变量间的回归关系。

12.2.4 结果分析

在 Stata 14.0 主界面的结果窗口可以看到如图 12.13~图 12.22 所示的分析结果。

1．对数据进行描述性分析的结果

图 12.13 是对数据进行描述性分析的结果。关于这一分析过程对于回归分析的重要意义已在前面章节中论述过，此处不再重复讲解。

在图 12.13 所示的分析结果中，可以得到很多信息，包括百分位数、4 个最小值、4 个最大值、平均值、标准差、偏度、峰度等。

```
. summarize age ratio,detail

                              age
-------------------------------------------------------------
      Percentiles      Smallest
 1%            1              1
 5%            1              2
10%            1              3       Obs                   7
25%            2              4       Sum of Wgt.           7

50%            4                      Mean                  4
                        Largest       Std. Dev.      2.160247
75%            6              4
90%            7              5       Variance       4.666667
95%            7              6       Skewness              0
99%            7              7       Kurtosis           1.75

                             ratio
-------------------------------------------------------------
      Percentiles      Smallest
 1%         56.7           56.7
 5%         56.7           75.9
10%         56.7           90.8       Obs                   7
25%         75.9           93.2       Sum of Wgt.           7

50%         93.2                      Mean           86.45714
                        Largest       Std. Dev.       14.9803
75%         96.3           93.2
90%         96.6           95.7       Variance       224.4095
95%         96.6           96.3       Skewness         -1.304
99%         96.6           96.6       Kurtosis       3.190059
```

图 12.13　对数据进行描述性分析

（1）百分位数（Percentiles）

可以看出变量 age 的第 1 个四分位数（25%）是 2，第 2 个四分位数（50%）是 4，第 3 个四分位数（75%）是 6；变量 ratio 的第 1 个四分位数（25%）是 75.9，第 2 个四分位数（50%）

是 93.2，第 3 个四分位数（75%）是 96.3。

（2）4 个最小值（Smallest）

变量 age 最小的 4 个数据值分别是 1、2、3、4。

变量 ratio 最小的 4 个数据值分别是 56.7、75.9、90.8、93.2。

（3）4 个最大值（Largest）

变量 age 最大的 4 个数据值分别是 4、5、6、7。

变量 ratio 最大的 4 个数据值分别是 93.2、95.7、96.3、96.6。

（4）平均值（Mean）和标准差（Std. Dev）

变量 age 的平均值为 4，标准差是 2.160247。

变量 ratio 的平均值为 86.45714，标准差是 14.9803。

（5）偏度（Skewness）和峰度（Kurtosis）

变量 age 的偏度为 0，为零偏度。

变量 ratio 的偏度为-1.304，为负偏度但不大。

变量 age 的峰度为 1.75，有一个比正态分布略短的尾巴。

变量 ratio 的峰度为 3.190059，有一个比正态分布略长的尾巴。

综上所述，数据的总体质量还是可以的，没有极端异常值，变量间的量纲差距、变量的偏度、峰度也是可以接受的，可以进行下一步的分析。

2．年龄和阴性率的线形图

图 12.14 是年龄和阴性率的线形图。

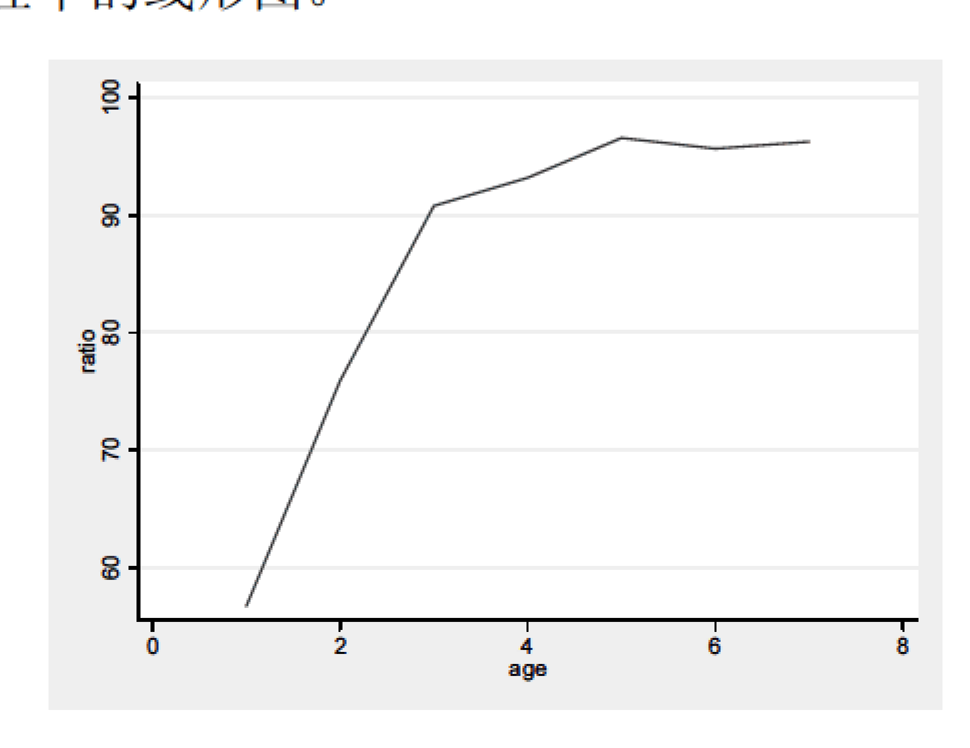


图 12.14 年龄和阴性率的线形图

从图 12.14 可以看出阴性率随着年龄的上升而上升，但是上升的速度越来越慢。

3．年龄和阴性率的散点图

图 12.15 是年龄和阴性率的散点图。

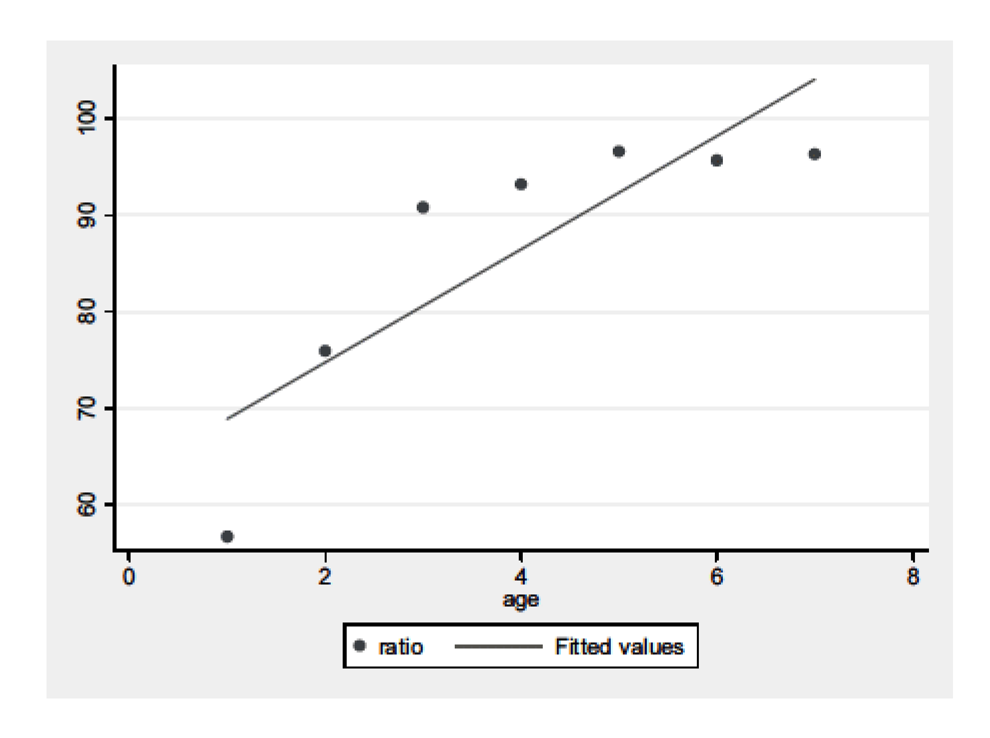


图 12.15　年龄和阴性率的散点图

从图 12.15 同样可以看出，阴性率随着年龄的上升而上升，但是上升的速度越来越慢，因此初步构想的模型包括线性、对数、二次、三次等。

4．对数据进行线性回归分析的结果

图 12.16 是对数据进行线性回归分析的结果。

```
. reg ratio age

      Source |       SS       df       MS              Number of obs =       7
-------------+------------------------------           F(  1,     5) =   12.55
       Model |  962.915714     1  962.915714           Prob > F      =  0.0165
    Residual |  383.541429     5  76.7082857           R-squared     =  0.7151
-------------+------------------------------           Adj R-squared =  0.6582
       Total |  1346.45714     6  224.409524           Root MSE      =  8.7583

------------------------------------------------------------------------------
       ratio |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   5.864286   1.655168     3.54   0.017     1.609541    10.11903
       _cons |         63   7.402137     8.51   0.000      43.9722     82.0278
------------------------------------------------------------------------------
```

图 12.16　对数据进行线性回归分析

从上述分析结果可以看出共有 7 个样本参与了分析，模型的 F 值(1, 5) = 12.55，P 值（Prob > F） = 0.0165，说明模型整体上是非常显著的。模型的可决系数（R-squared） = 0.7151，模型修正的可决系数（Adj R-squared） = 0.6582，说明模型的解释能力还是差强人意的。

变量 age 的系数标准误是 1.655168，t 值为 3.54，P 值为 0.017，系数是非常显著的，95%的置信区间为[1.609541, 10.11903]。常数项的系数标准误是 7.402137，t 值为 8.51，P 值为 0.000，系数也是非常显著的，95%的置信区间为[43.9722，82.0278]。

模型的回归方程是：

ratio = 5.864286* age + 63

从上面的分析可以看出线性模型的整体显著性和系数显著性尚可，但模型的整体解释能力有较大提升空间。

5．对数据进行对数变换线性回归分析的结果

选择“Data” | “Data Editor” | “Data Editor(Browse)” 命令，进入数据查看界面，可以看到如图 12.17 所示的 lnage 数据。

图 12.18 是对数据进行对数变换线性回归分析的结果。

	age	ratio	age2	age3	lnage
1	1	56.7	1	1	0
2	2	75.9	4	8	.6931472
3	3	90.8	9	27	1.098612
4	4	93.2	16	64	1.386294
5	5	96.6	25	125	1.609438
6	6	95.7	36	216	1.791759
7	7	96.3	49	343	1.94591

图 12.17　数据查看界面

```
. reg ratio lnage

      Source |       SS       df       MS              Number of obs =       7
-------------+------------------------------           F(  1,      5) =   53.00
       Model |  1230.38048     1  1230.38048           Prob > F      =  0.0008
    Residual |   116.07666     5   23.215332           R-squared     =  0.9138
-------------+------------------------------           Adj R-squared =  0.8965
       Total |  1346.45714     6  224.409524           Root MSE      =  4.8182

------------------------------------------------------------------------------
       ratio |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lnage |   20.91074   2.872349     7.28   0.001     13.52713    28.29435
       _cons |   60.99036    3.94382    15.46   0.000     50.85245    71.12828
------------------------------------------------------------------------------
```

图 12.18　对数据进行对数变换线性回归分析

从上述分析结果中可以看出模型的 F 值(1, 5)升为 53，P 值（Prob > F） 升为 0.0008，说明模型整体显著程度继续上升。模型的可决系数（R-squared）= 0.9138，模型修正的可决系数（Adj R-squared）= 0.8965，说明模型的解释能力大幅度提升。

变量 lnage 的系数标准误是 2.872349，t 值为 7.28，P 值为 0.001，系数是非常显著的，95%的置信区间为[13.52713, 28.29435]。常数项的系数标准误是 3.94382，t 值为 15.46，P 值为 0.000，系数也是非常显著的，95%的置信区间为[50.85245，71.12828]。

模型的回归方程是：

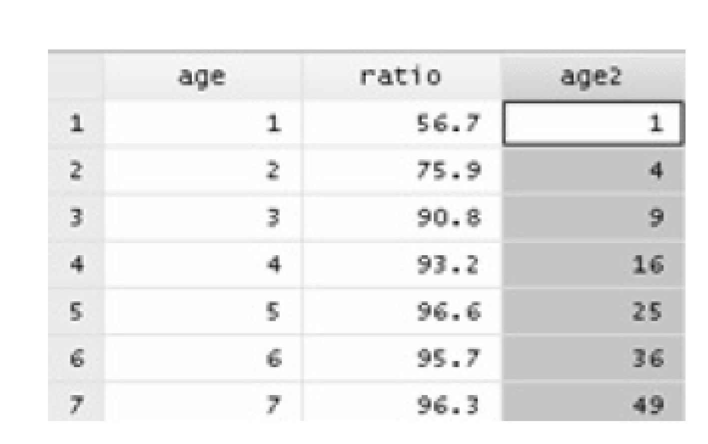
ratio = 20.91074* lnage + 60.99036

从上面的分析可以看出对数模型的整体显著性和系数显著性较线性模型虽略有升高，但对模型的整体解释能力却有了较大提升。

6．对数据进行二次变换线性回归分析的结果

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 12.19 所示的 age2 数据。

图 12.20 是对数据进行二次变换线性回归分析的结果。



	age	ratio	age2
1	1	56.7	1
2	2	75.9	4
3	3	90.8	9
4	4	93.2	16
5	5	96.6	25
6	6	95.7	36
7	7	96.3	49

图 12.19　数据查看界面

```
. reg  ratio age2 age

      Source |       SS       df       MS              Number of obs =       7
-------------+------------------------------           F(  2,      4) =   66.19
       Model |  1306.96333     2  653.481667           Prob > F      =  0.0009
    Residual |  39.4938095     4  9.87345238           R-squared     =  0.9707
-------------+------------------------------           Adj R-squared =  0.9560
       Total |  1346.45714     6  224.409524           Root MSE      =  3.1422

------------------------------------------------------------------------------
       ratio |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        age2 |   -2.02381   .3428427    -5.90   0.004    -2.975693   -1.071926
         age |   22.05476   2.806288     7.86   0.001     14.26326    29.84627
       _cons |   38.71429   4.896773     7.91   0.001     25.11866    52.30991
------------------------------------------------------------------------------
```

图 12.20　对数据进行二次变换线性回归分析

从上述分析结果中可以看出模型的 F 值(2,4)上升为 66.19，P 值（Prob > F）为 0.0009，说明模型整体显著程度依旧非常好。模型的可决系数（R-squared）= 0.9707，模型修正的可决系数（Adj R-squared）= 0.9560，说明模型的解释能力又有小幅度提升。

变量 age2 的系数标准误是 0.3428427，t 值为-5.90，P 值为 0.004，系数是非常显著的，95%

的置信区间为[-2.975693, -1.071926]。变量 age 的系数标准误是 2.806288，t 值为 7.86，P 值为 0.001，系数是非常显著的，95%的置信区间为[14.26326, 29.84627]。常数项的系数标准误是 4.896773，t 值为 7.91，P 值为 0.001，系数也是非常显著的，95%的置信区间为[25.11866, 52.30991]。

模型的回归方程是：

ratio = -2.02381* age2 + 22.05476* age+38.71429

从上面的分析可以看出二次模型在保持整体显著性和系数显著性的同时，实现了模型整体解释能力的小幅度提升。

7．对数据进行三次变换线性回归分析的结果

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 12.21 所示的 age3 数据。

图 12.22 是对数据进行三次变换线性回归分析的结果。

	age	ratio	age2	age3
1	1	56.7	1	1
2	2	75.9	4	8
3	3	90.8	9	27
4	4	93.2	16	64
5	5	96.6	25	125
6	6	95.7	36	216
7	7	96.3	49	343

图 12.21 数据查看界面

```
. reg ratio age3 age2 age

      Source |       SS       df       MS              Number of obs =       7
-------------+------------------------------           F(  3,     3) =  196.22
       Model |     1339.63     3  446.543333           Prob > F      =  0.0006
    Residual |  6.82714286     3  2.27571429           R-squared     =  0.9949
-------------+------------------------------           Adj R-squared =  0.9899
       Total |  1346.45714     6  224.409524           Root MSE      =  1.5085

------------------------------------------------------------------------------
       ratio |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        age3 |   .3888889   .1026436     3.79   0.032     .0622311    .7155467
        age2 |  -6.690476   1.242672    -5.38   0.013    -10.64521   -2.735738
         age |   37.99921   4.418788     8.60   0.003     23.93665    52.06176
       _cons |   24.71429   4.379614     5.64   0.011      10.7764    38.65217
------------------------------------------------------------------------------
```

图 12.22 对数据进行三次变换线性回归分析

从上述分析结果中可以看出模型的 F 值(3,3)上升为 196.22，P 值（Prob > F）为 0.0006，说明模型整体显著程度继续上升。模型的可决系数（R-squared）= 0.9949，模型修正的可决系数（Adj R-squared）= 0.9899，说明模型的解释能力又有小幅度提升，接近完美。

变量 age3 的系数标准误是 0.1026436，t 值为 3.79，P 值为 0.032，系数是非常显著的，95%的置信区间为[0.0622311, 0.7155467]。变量 age2 的系数标准误是 1.242672，t 值为-5.38，P 值为 0.013，系数是非常显著的，95%的置信区间为[-10.64521, -2.735738]。变量 age 的系数标准误是 4.418788，t 值为 8.60，P 值为 0.003，系数是非常显著的，95%的置信区间为[23.93665, 52.06176]。常数项的系数标准误是 4.379614，t 值为 5.64，P 值为 0.011，系数也是非常显著的，95%的置信区间为[10.7764,38.65217]。

模型的回归方程是：

ratio =0.3888889* age3 -6.690476* age2 + 37.99921* age+24.71429

从上面的分析可以看出三次模型在保持整体显著性和系数显著性的同时，又实现了模型整体解释能力的小幅度提升，使模型接近完美。

12.2.5　案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

下面采用前面介绍过的 sw regress 命令选择回归模型自变量。

可以定义年龄 age、年龄的二次方 age2、年龄的三次方 age3、年龄的四次方 age4、年龄的五次方 age5 自变量，并设定显著性水平为 0.05，操作命令如下：

```
sw regress  ratio age age2 age3 age4 age5,pr(0.05)
```

在命令窗口输入命令并按回车键进行确认，结果如图 12.23 所示。

```
. sw regress   ratio age age2 age3 age4 age5,pr(0.05)
                      begin with full model
p = 0.9806 >= 0.0500  removing age
p = 0.1301 >= 0.0500  removing age5

      Source |       SS           df       MS      Number of obs   =         7
-------------+----------------------------------   F(3, 3)         =     59.99
       Model |  1324.38121         3  441.460403   Prob > F        =    0.0035
    Residual |  22.0759343         3  7.35864476   R-squared       =    0.9836
-------------+----------------------------------   Adj R-squared   =    0.9672
       Total |  1346.45714         6  224.409524   Root MSE        =    2.7127

------------------------------------------------------------------------------
       ratio |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        age4 |   .1907789   .0418335     4.56   0.020     .0576461    .3239117
        age2 |   10.84041   1.591905     6.81   0.006     5.774258    15.90656
        age3 |  -2.746933   .5064152    -5.42   0.012    -4.358572   -1.135294
       _cons |   49.86513   3.332115    14.97   0.001     39.26086    60.46941
------------------------------------------------------------------------------
```

图 12.23　分析结果图

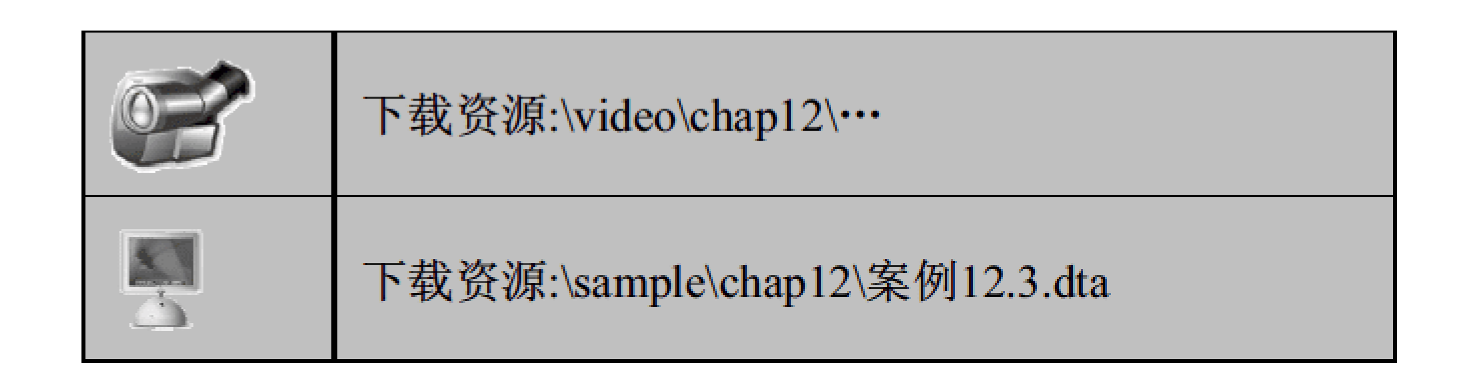
至于本结果的详细解读与前面重复，限于篇幅，这里不再赘述。

12.3　实例三——非线性回归分析

12.3.1　非线性回归分析的功能与意义

上节讲述的转换变量回归分析从本质上讲仍属于一种线性回归分析方法，而实际问题往往会更复杂，使用转换变量回归分析方法便无法做出准确的分析，这时候就需要用到 Stata 的非线性回归分析。非线性回归分析是一种功能更强大的处理非线性问题的方法，可以使用户自定义任意形式的函数，从而更加准确地描述变量之间的关系。

12.3.2　相关数据来源



	下载资源:\video\chap12\…
	下载资源:\sample\chap12\案例12.3.dta

【例 12.3】某著名总裁培训班的讲师想要建立一个回归模型，对参与培训的企业高管毕业后的长期表现情况进行预测。自变量是高管的培训天数，因变量是高管毕业后的长期表现指数，指数越大，表现越好。表 12.3 给出了相关数据，试用非线性回归方法拟合模型。

表 12.3　15 名高管的培训天数（x）与长期表现指数（y）

编号	培训天数	长期表现指数
1	2	53
2	65	6
3	52	11
4	60	4
5	14	34
6	53	8
7	10	36
8	26	19
9	19	26
10	31	16
11	38	13
12	45	8
13	34	19
14	7	45
15	5	51

12.3.3　Stata 分析过程

在用 Stata 进行分析之前，要把数据录入到 Stata 中。本例中有两个变量，分别是培训天数和长期表现指数。把培训天数变量设定为 x，把长期表现指数变量设定为 y，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作已在第 1 章中有过详细讲述。录入完成后数据如图 12.24 所示。

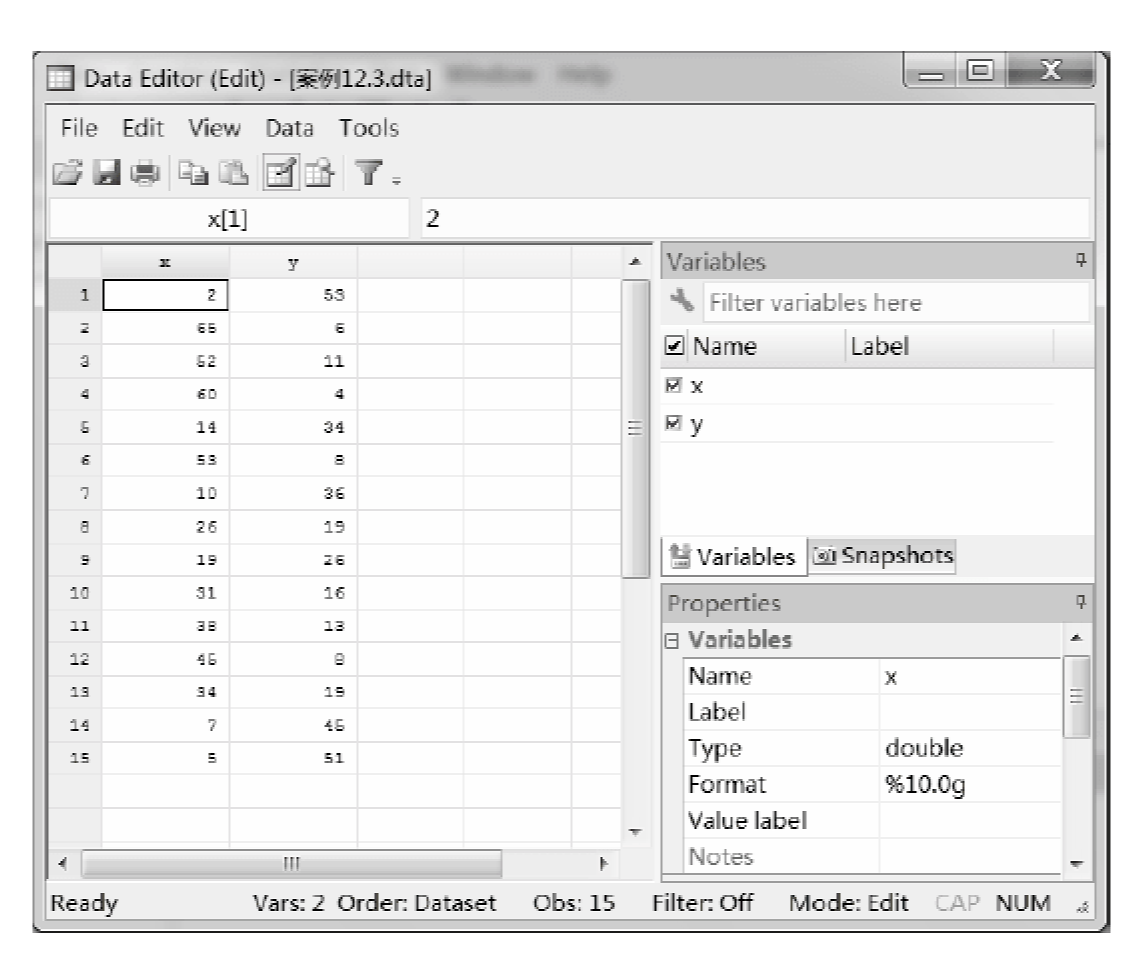


图 12.24　案例 12.3 数据

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入如下命令。

- summarize y x,detail：本命令的含义是对长期表现指数和培训天数进行描述性分析，简要探索数据特征，从整体上对数据有一个清晰直观的把握。
- twoway line y x：本命令旨在通过绘制长期表现指数和培训天数的线形图，从整体上对数据有一个清晰直观的把握。
- graph twoway scatter y x || lfit y x：本命令旨在通过绘制长期表现指数和培训天数的散点图，从整体上对数据有一个清晰直观的把握。
- reg y x：本命令旨在构建线性模型，以长期表现指数为因变量，以培训天数为自变量，进行最小二乘回归分析，探索变量间的回归关系。
- nl (y = exp({a}+{b}*x))：本命令旨在以长期表现指数为因变量，以培训天数为自变量，构建非线性模型 y = exp({a}+{b}*x)，进行非线性回归分析。
- vce：本命令旨在估计系数 a 和 b 的方差-协方差矩阵。
- predict yhat：本命令旨在获得因变量的拟合值。
- predict e,resid：本命令旨在获得回归模型的估计残差。

03 设置完毕后，按键盘上的回车键，等待输出结果。

12.3.4 结果分析

在 Stata 14.0 主界面的结果窗口可以看到如图 12.25~图 12.32 所示的分析结果。

1．对数据进行描述性分析的结果

图 12.25 是对数据进行描述性分析的结果。关于这一分析过程对于回归分析的重要意义在前面章节中已经论述过，此处不再重复讲解。

```
. summarize y x,detail

                              y
-------------------------------------------------------------
      Percentiles      Smallest
 1%            4              4
 5%            4              6
10%            6              8       Obs                  15
25%            8              8       Sum of Wgt.          15

50%           19                      Mean           23.26667
                        Largest       Std. Dev.      16.67105
75%           36             36
90%           51             45       Variance       277.9238
95%           53             51       Skewness        .611507
99%           53             53       Kurtosis       1.909912

                              x
-------------------------------------------------------------
      Percentiles      Smallest
 1%            2              2
 5%            2              5
10%            5              7       Obs                  15
25%           10             10       Sum of Wgt.          15

50%           31                      Mean           30.73333
                        Largest       Std. Dev.      20.98798
75%           52             52
90%           60             53       Variance       440.4952
95%           65             60       Skewness       .1586165
99%           65             65       Kurtosis       1.706699
```

图 12.25 分析结果图

在如图 12.25 所示的分析结果中，可以得到很多信息，包括百分位数、4 个最小值、4 个最大值、平均值、标准值、偏度、峰度等。

（1）百分位数（Percentiles）

可以看出变量 y 的第 1 个四分位数（25%）是 8，第 2 个四分位数（50%）是 19，第 3 个四分位数（75%）是 36；变量 x 的第 1 个四分位数（25%）是 10，第 2 个四分位数（50%）是 31，第 3 个四分位数（75%）是 52。

（2）4 个最小值（Smallest）

变量 y 最小的 4 个数据值分别是 4、6、8、8。

变量 x 最小的 4 个数据值分别是 2、5、7、10。

（3）4 个最大值（Largest）

变量 y 最大的 4 个数据值分别是 36、45、51、53。

变量 x 最大的 4 个数据值分别是 52、53、60、65。

（4）平均值（Mean）和标准差（Std. Dev）

变量 y 的平均值为 23.26667，标准差是 16.67105。

变量 x 的平均值为 30.73333，标准差是 20.98798。

（5）偏度（Skewness）和峰度（Kurtosis）

变量 y 的偏度为 0.611507，为正偏度但不大。

变量 x 的偏度为 0.1586165，为正偏度但不大。

变量 y 的峰度为 1.989912，有一个比正态分布略短的尾巴。

变量 x 的峰度为 1.706699，有一个比正态分布略短的尾巴。

综上所述，数据的总体质量还是可以的，没有极端异常值，变量间的量纲差距、变量的偏度、峰度也是可以接受的，可以进入下一步的分析。

图 12.26 是长期表现指数和培训天数的线形图。

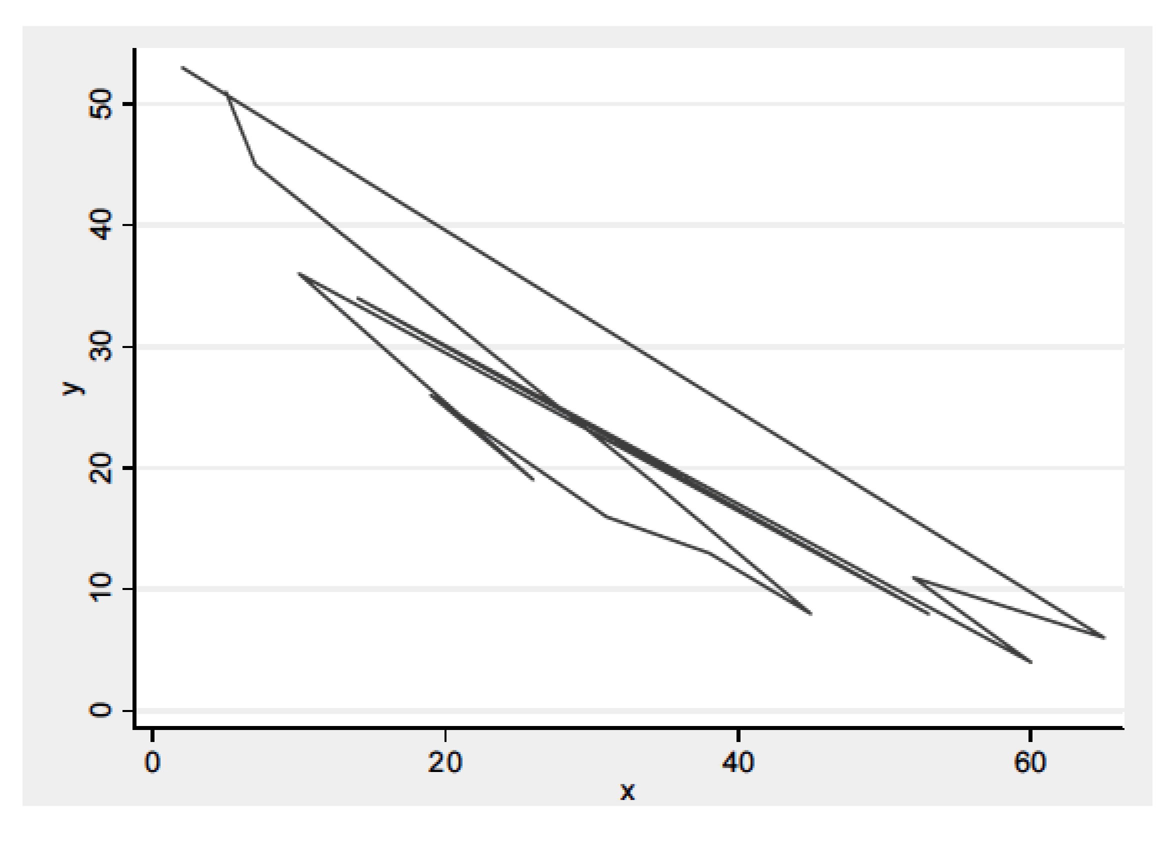


图 12.26　长期表现指数和培训天数的线形图

从图 12.26 可以看出长期表现指数随着培训天数的上升而上升，但是上升的逐渐程度不明朗。

图 12.27 是长期表现指数和培训天数的散点图。

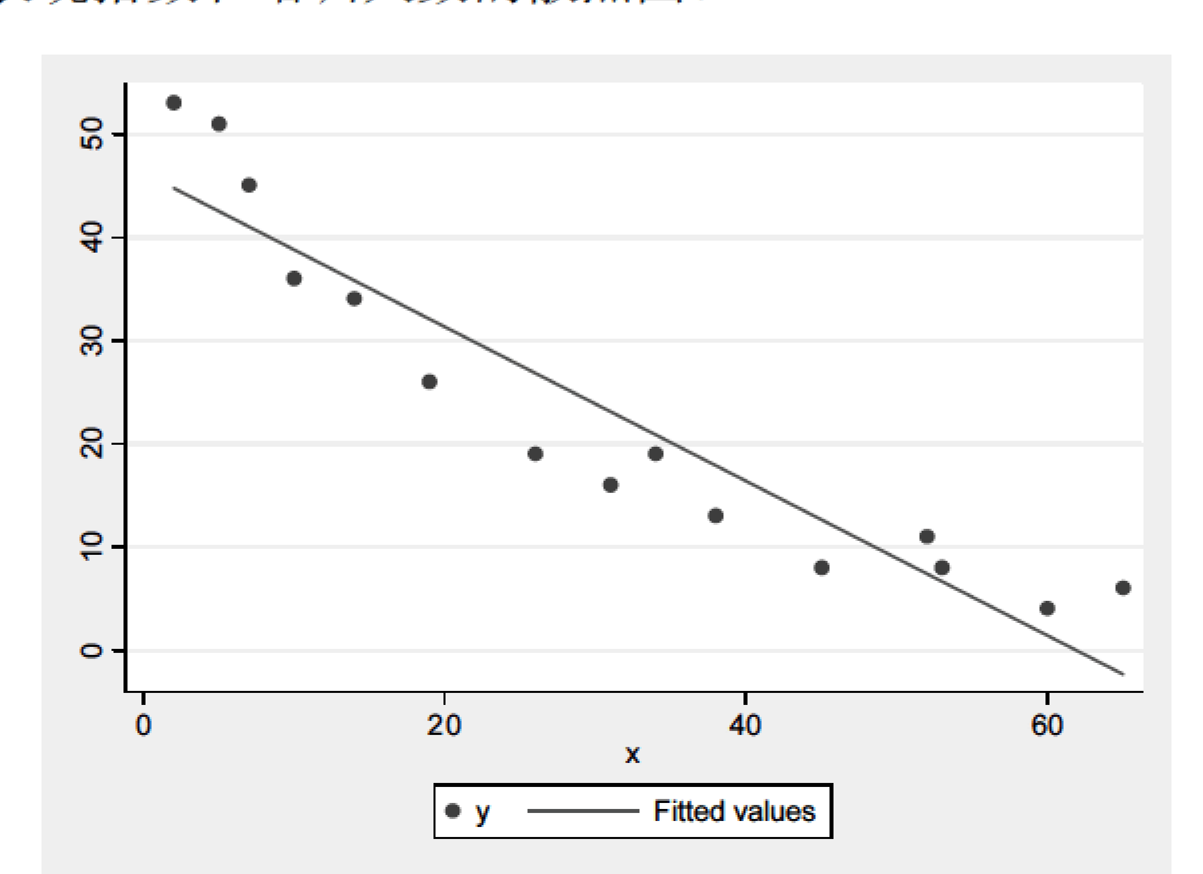


图 12.27 长期表现指数和培训天数的散点图

从图 12.27 同样可以看出长期表现指数随着培训天数的上升而上升，但是上升的逐渐程度不明朗。

图 12.28 是对数据进行线性回归分析的结果。、

```
. regress y x

      Source |       SS       df       MS              Number of obs =      15
-------------+------------------------------           F(  1,    13) =   98.45
       Model |  3437.07334     1  3437.07334           Prob > F      =  0.0000
    Residual |  453.859995    13  34.9123073           R-squared     =  0.8834
-------------+------------------------------           Adj R-squared =  0.8744
       Total |  3890.93333    14   277.92381           Root MSE      =  5.9087

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |  -.7465515    .075241    -9.92   0.000    -.9090998   -.5840032
       _cons |   46.21068   2.770327    16.68   0.000     40.22575    52.19561
------------------------------------------------------------------------------
```

图 12.28 对数据进行线性回归分析

从上述分析结果中可以得到很多信息。可以看出共有 15 个样本参与了分析，模型的 F 值 (1, 13) = 98.45，P 值（Prob > F）= 0.0000，说明模型整体上是非常显著的。模型的可决系数（R-squared）为 0.8834，模型修正的可决系数（Adj R-squared）为 0.8744，说明模型的解释能力还是差强人意的。

变量 x 的系数标准误是 0.075241，t 值为-9.92，P 值为 0.000，系数是非常显著的，95%的置信区间为[-0.9090998, -0.5840032]。常数项的系数标准误是 2.770327，t 值为 16.68，P 值为 0.000，系数也是非常显著的，95%的置信区间为[40.22575,52.19561]。

模型的回归方程是：

y= -0.7465515 *x + 46.21068

从上面的分析可以看出线性模型的整体显著性和系数显著性尚可，但模型的整体解释能力有较大提升空间。

图 12.29 是对数据进行非线性回归分析的结果。

```
. nl (y = exp({a}+{b}*x))
(obs = 15)

Iteration 0:  residual SS =  6452.563
Iteration 1:  residual SS =  181.1452
Iteration 2:  residual SS =  66.15499
Iteration 3:  residual SS =  64.57034
Iteration 4:  residual SS =  64.56715
Iteration 5:  residual SS =  64.56715
Iteration 6:  residual SS =  64.56715

      Source |       SS       df       MS
-------------+------------------------------         Number of obs =         15
       Model |  11946.4329     2  5973.21643         R-squared     =     0.9946
    Residual |   64.567146    13  4.96670354         Adj R-squared =     0.9938
-------------+------------------------------         Root MSE      =    2.22861
       Total |       12011    15  800.733333         Res. dev.     =   64.46299

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          /a |   4.063108   .0288334   140.92   0.000     4.000817    4.125399
          /b |  -.0392997   .0019524   -20.13   0.000    -.0435175   -.0350819
------------------------------------------------------------------------------
```

图 12.29　对数据进行非线性回归分析

从上述分析结果中可以得到很多信息。可以看出模型的可决系数（R-squared）大幅上升为 0.9946，模型修正的可决系数（Adj R-squared）为 0.9938，模型的解释能力几乎达到完美状态。

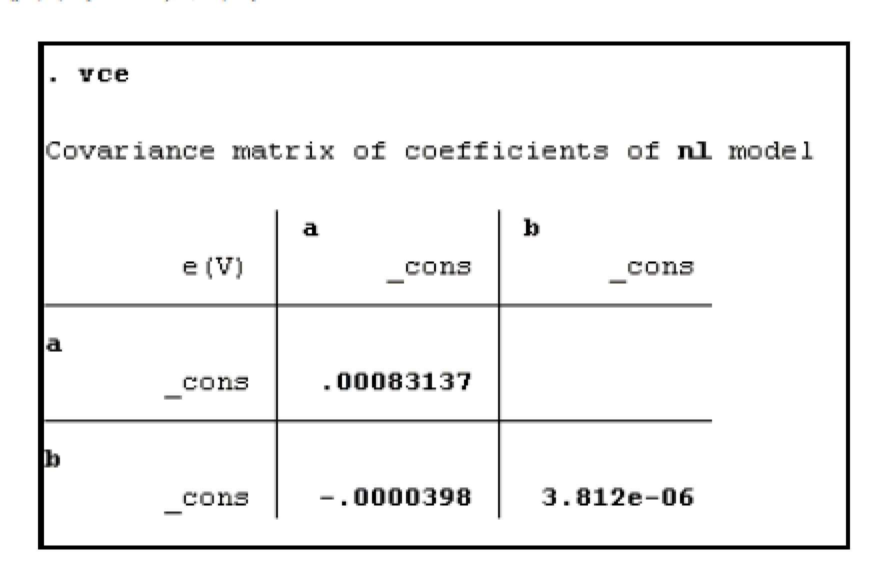
系数 a 的系数标准误是 0.0288334，t 值为 140.92，P 值为 0.000，系数是非常显著的，95%的置信区间为[4.000817, 4.125399]。系数 b 的系数标准误是 0.0019524，t 值为-20.13，P 值为 0.000，系数也是非常显著的，95%的置信区间为[-0.0435175,-0.0350819]。

模型的回归方程是：

y=EXP(4.063108 -0.0392997*x)

从上面的分析可以看出非线性回归模型在保持整体显著性和系数显著性较线性模型很高的基础上，实现了模型的整体解释能力的较大提升。

图 12.30 是系数的方差-协方差矩阵。



. vce

Covariance matrix of coefficients of nl model

e(V)	a _cons	b _cons
a _cons	.00083137	
b _cons	-.0000398	3.812e-06

图 12.30　系统的方差－协方差矩阵

从图 12.30 中可以看出，系数间的方差与协方差都不是很大，有些甚至微不足道。

图 12.31 是对因变量的拟合值的预测。

关于因变量预测拟合值的意义我们在前面章节中已经论述了，此处不再重复讲解。

图 12.32 是回归分析得到的残差序列。

	x	y	yhat
1	2	53	53.75887
2	65	6	4.520528
3	52	11	7.534743
4	60	4	5.502086
5	14	34	33.54583
6	53	8	7.244373
7	10	36	39.25627
8	26	19	20.93278
9	19	26	27.56134
10	31	16	17.19843
11	38	13	13.06217
12	45	8	9.92069
13	34	19	15.28572
14	7	45	44.16843
15	5	51	47.78012

图 12.31　因变量的拟合值预测

	x	y	yhat	e
1	2	53	53.75887	-.7588684
2	65	6	4.520528	1.479472
3	52	11	7.534743	3.465257
4	60	4	5.502086	-1.502086
5	14	34	33.54583	.454172
6	53	8	7.244373	.7556269
7	10	36	39.25627	-3.256272
8	26	19	20.93278	-1.93278
9	19	26	27.56134	-1.561341
10	31	16	17.19843	-1.198428
11	38	13	13.06217	-.0621694
12	45	8	9.92069	-1.92069
13	34	19	15.28572	3.714282
14	7	45	44.16843	.8315712
15	5	51	47.78012	3.219882

图 12.32　残差序列

关于残差序列的意义我们在前面章节中已经论述了，此处不再重复讲解。

12.3.5　案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

1．延伸 1：设定非线性回归模型中被估计参数的初始值

例如，本例中我们把系数 a 的起始值设定为 4，把系数 b 的初始值设定为-0.04，那么操作命令可以相应地修改为：

```
nl (y = exp({a}+{b}*x)), initial(a 4 b -0.04)
```

在命令窗口输入命令并按回车键进行确认，结果如图 12.33 所示。

```
. nl (y = exp({a}+{b}*x)), initial(a 4 b -0.04)
(obs = 15)

Iteration 0:  residual SS =  64.64718
Iteration 1:  residual SS =  64.56715
Iteration 2:  residual SS =  64.56715
Iteration 3:  residual SS =  64.56715

      Source |      SS           df       MS
-------------+----------------------------------    Number of obs =         15
       Model |  11946.4329        2  5973.21643    R-squared     =     0.9946
    Residual |   64.567146       13  4.96670354    Adj R-squared =     0.9938
-------------+----------------------------------    Root MSE      =    2.22861
       Total |       12011       15  800.733333    Res. dev.     =   64.46299

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          /a |   4.063108   .0288334   140.92   0.000     4.000817    4.125399
          /b |  -.0392997   .0019524   -20.13   0.000    -.0435175   -.0350819
------------------------------------------------------------------------------
```

图 12.33　设定非线性回归模型中被估计参数的初始值

从上面的分析结果中可以看出由于初始参数值的设定减少了迭代次数，提高了系统运行效率，但结果与前面是一致的，对本结果的详细解读限于篇幅不再赘述。

2．延伸 2：采用稳健的标准差进行非线性回归估计

与线性回归类似，非线性回归也可以允许稳健标准差选择项的存在，例如本例如果使用

稳健的标准差，那么操作命令就是：

```
nl (y = exp({a}+{b}*x)),robust
```

在命令窗口输入命令并按回车键进行确认，结果如图 12.34 所示。

上面的分析结果与没有使用稳健标准差进行回归时大同小异，对本结果的详细解读限于篇幅不再赘述。

```
. nl (y = exp({a}+{b}*x)),robust
(obs = 15)

Iteration 0:  residual SS =  6452.563
Iteration 1:  residual SS =  181.1452
Iteration 2:  residual SS =  66.15499
Iteration 3:  residual SS =  64.57034
Iteration 4:  residual SS =  64.56715
Iteration 5:  residual SS =  64.56715
Iteration 6:  residual SS =  64.56715

Nonlinear regression                              Number of obs =         15
                                                  R-squared     =     0.9946
                                                  Adj R-squared =     0.9938
                                                  Root MSE      =    2.22861
                                                  Res. dev.     =   64.46299

------------------------------------------------------------------------------
             |               Robust
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          /a |   4.063108   .0279161   145.55   0.000     4.002799    4.123417
          /b |  -.0392997   .0018994   -20.69   0.000    -.0434031   -.0351962
------------------------------------------------------------------------------
```

图 12.34　采用稳健的标准差进行非线性回归估计

3. 延伸 3：采用系统默认快捷函数进行非线性回归

由于很多非线性函数常常被用到，因此 Stata 将这些函数进行了内置，用户在使用时可以轻松地使用简易命令调出，而不必输入复杂的模型方程形式。Stata 内置非线性函数命令缩写与函数形式如表 12.4 所示。

表 12.4　Stata 内置非线性函数命令缩写与函数形式

非线性函数命令缩写	非线性函数形式
exp2	y =b1*b2^x
exp3	y = b0 + b1*b2^x
exp2a	y = b1*(1-b2^x)
log3	y = b1/(1 + exp(-b2*(x-b3)))
log4	y = b0 + b1/(1 + exp(-b2*(x-b3)))
gom3	y =b1*exp(-exp(-b2*(x-b3)))
gom4	y = b0 + b1*exp(-exp(-b2*(x-b3)))

例如，在本例中如果我们设定非线性模型回归形式为：y =b1*b2^x，那么操作命令就是：

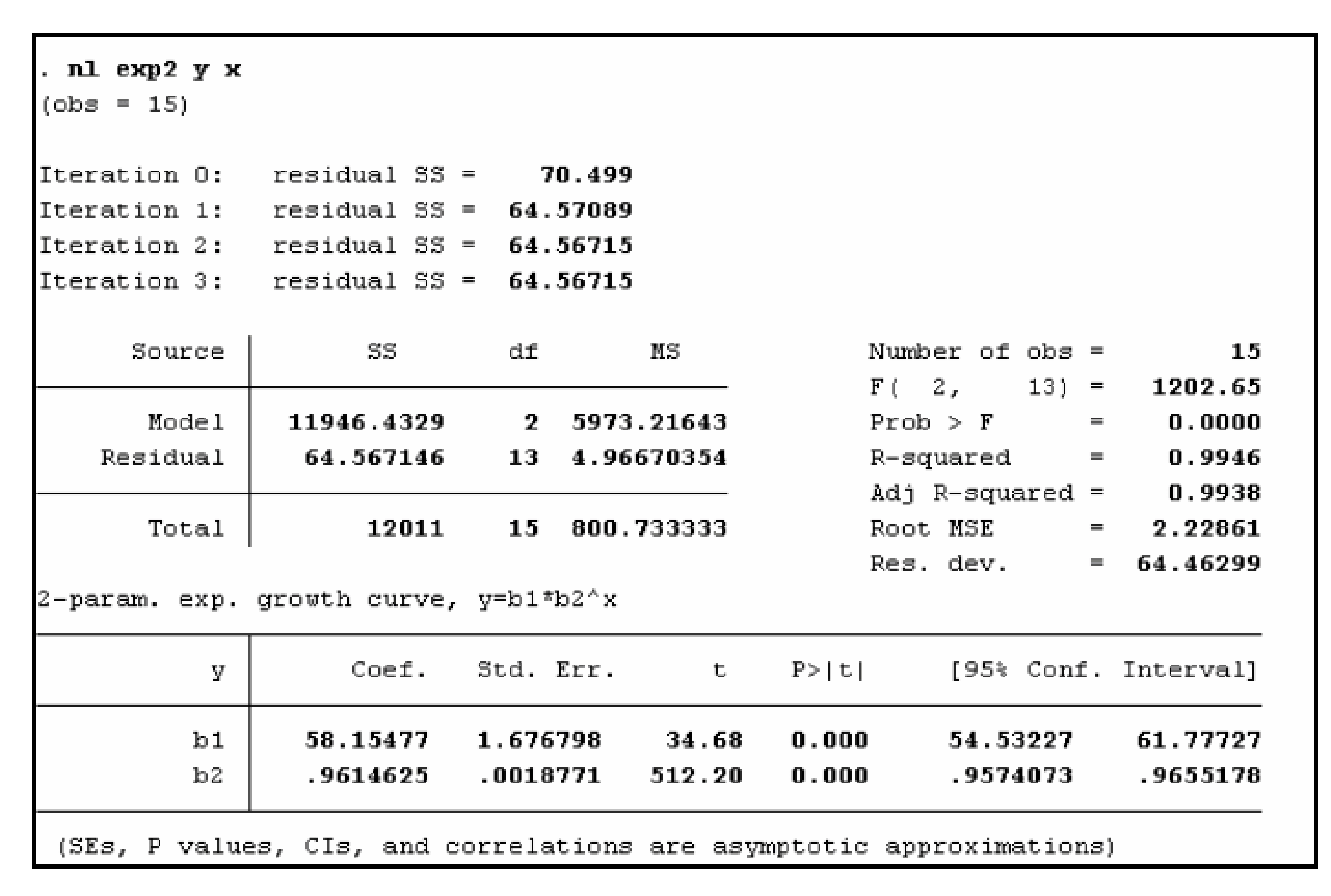
```
nl exp2 y x
```

在命令窗口输入命令并按回车键进行确认，结果如图 12.35 所示。

对该模型结果的详细解读限于篇幅不再赘述。我们得到的非线性回归方程是：

y = 58.15477 *0.9614625 ^x

模型的解释能力和显著性都非常好。



```
. nl exp2 y x
(obs = 15)

Iteration 0:   residual SS =     70.499
Iteration 1:   residual SS =   64.57089
Iteration 2:   residual SS =   64.56715
Iteration 3:   residual SS =   64.56715

      Source |       SS       df       MS            Number of obs =        15
-------------+------------------------------         F(  2,     13) =   1202.65
       Model |  11946.4329     2  5973.21643         Prob > F      =    0.0000
    Residual |   64.567146    13  4.96670354         R-squared     =    0.9946
-------------+------------------------------         Adj R-squared =    0.9938
       Total |       12011    15  800.733333         Root MSE      =   2.22861
                                                     Res. dev.     =  64.46299
2-param. exp. growth curve, y=b1*b2^x
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          b1 |   58.15477   1.676798    34.68   0.000     54.53227    61.77727
          b2 |   .9614625   .0018771   512.20   0.000     .9574073    .9655178
------------------------------------------------------------------------------
  (SEs, P values, CIs, and correlations are asymptotic approximations)
```

图 12.35　采用系统默认快捷函数进行非线性回归

12.4　本章习题

（1）某两家足球俱乐部的部分球员历年进球数如表 12.5 所示，请用非参数回归方法研究年份和绩效考核系数两个变量之间的关系。

表 12.5　某两家足球俱乐部的部分球员历年进球数

所属俱乐部	年份	平均进球数
A俱乐部	2000	1.8
A俱乐部	2000	2
A俱乐部	2000	1.9
A俱乐部	2001	1.7
A俱乐部	2001	1.6
…	…	…
B俱乐部	2010	1.49
B俱乐部	2010	1.69
B俱乐部	2010	1.92

（2）某著名总裁培训班的讲师想建立一个回归模型，对参与培训的企业高管毕业后的长期表现情况进行预测。自变量是高管的培训天数，因变量是高管毕业后的长期表现指数，指数越大，表现越好。表 12.6 给出了相关数据，试用转换变量回归分析方法拟合曲线。

表 12.6　15 名高管的培训天数（x）与长期表现指数（y）

编号	培训天数	长期表现指数
1	2	53
2	65	6
3	52	11
4	60	4
5	14	34

（续表）

编号	培训天数	长期表现指数
6	53	8
7	10	36
8	26	19
9	19	26
10	31	16
11	38	13
12	45	8
13	34	19
14	7	45
15	5	51

（3）研究发现，锡克氏试验阴性率随着儿童年龄的增长而升高。查得山东省某地 1~7 岁儿童的资料如表 12.7 所示，试用非线性回归方法拟合模型。

表 12.7　儿童锡克氏试验阴性率

年龄/岁	阴性率/%
1	56.7
2	75.9
3	90.8
4	93.2
5	96.6
6	95.7
7	96.3

第 13 章　Stata Logistic 回归分析

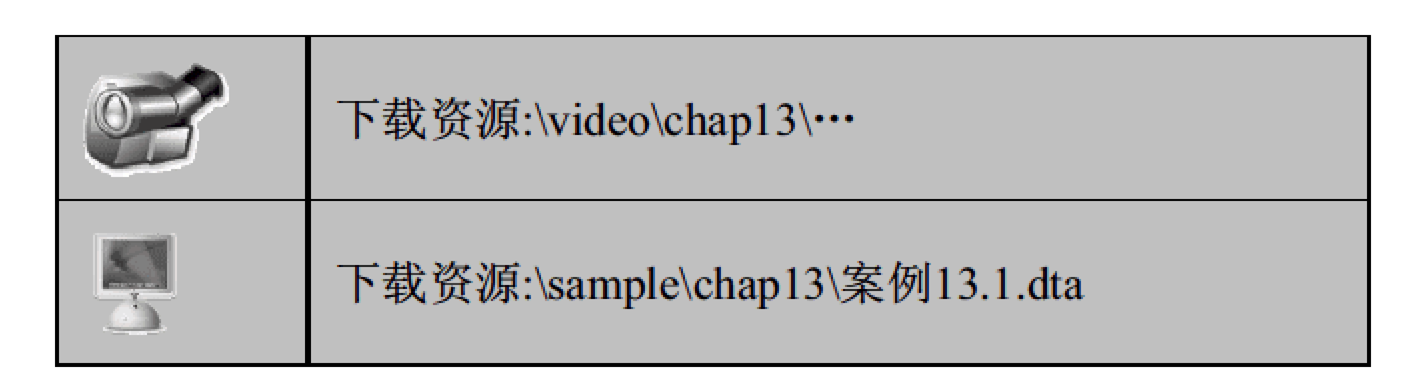
前面我们讲述的回归分析方法都要求因变量是连续变量，但很多情况下因变量是离散的，而非连续的。例如，公司招聘人才时根据对应聘人员的特征做出录用或者不录用的评价、毕业学生对职业的选择等。这时就需要用到我们本章介绍的 Logistic 回归分析。根据因变量的离散特征，常用的 Logistic 回归分析方法有 3 种，包括二元 Logistic 回归分析、多元 Logistic 回归分析以及有序 Logistic 回归分析等。下面我们就以实例的方式一一介绍这几种方法在 Stata 中的应用。

13.1　实例一——二元Logistic回归分析

13.1.1　二元 logistic 回归分析的功能与意义

我们经常会遇到因变量只有两种取值的情况，例如是否患病、是否下雨等，这时一般的线性回归分析将无法准确刻画变量之间的因果关系，需要用其他的回归分析方法来进行拟合模型。Stata 的二项分类 Logistic 回归便是一种简便的处理二分类因变量问题的分析方法。

13.1.2　相关数据来源



【例 13.1】表 13.1 给出了 20 名肾癌患者的相关数据。试用二项分类 Logistic 回归方法分析患者肾细胞癌转移情况（有转移 y=1、无转移 y=0）与患者年龄、肾细胞癌血管内皮生长因子（其阳性表示由低到高共 3 个等级）、肾癌细胞核组织学分级（由低到高共 4 级）、肾细胞癌组织内微血管数、肾细胞癌分期（由低到高共 4 期 ）之间的关系。

表 13.1　20 名肾癌患者的相关数据

编号	肾细胞癌转移情况	年龄/岁	肾细胞癌血管内皮生长因子	肾癌细胞核组织学分级	肾细胞癌组织内微血管数/个/μL	肾细胞癌分期
1	0	60	3	3	46	1
2	1	35	2	2	60	2
3	1	64	1	1	146	3
4	0	67	2	3	100	2
5	0	54	3	4	92	3

（续表）

编号	肾细胞癌转移情况	年龄/岁	肾细胞癌血管内皮生长因子	肾癌细胞核组织学分级	肾细胞癌组织内微血管数/个/μL	肾细胞癌分期
6	0	57	3	3	98	2
7	1	40	1	2	70	1
8	0	41	2	4	202	4
9	0	51	1	1	76	1
10	1	57	3	1	70	2
11	0	66	2	3	123	1
12	1	30	3	4	89	3
13	0	53	1	1	59	1
14	0	34	3	2	49	2
15	1	38	1	4	35	3
16	0	41	1	2	67	1
17	0	16	1	3	134	1
18	1	34	3	2	116	3
19	1	46	1	2	51	3
20	0	72	3	4	180	2

13.1.3 Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有 6 个变量，分别是肾细胞癌转移情况、年龄、肾细胞癌血管内皮生长因子、肾癌细胞核组织学分级、肾细胞癌组织内微血管数和肾细胞癌分期。我们把这 6 个变量分别定义为 V1、V2、V3、V4、V5、V6。变量类型及长度采取系统默认方式，然后录入相关数据。相关操作我们在第 1 章中已有详细讲述。录入完成后数据如图 13.1 所示。

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入如下命令。

- list V1-V6：本命令的含义是对 6 个变量所包含的样本数据进行一一展示，以便简单直观地观测出数据的具体特征，为深入分析做好必要准备。
- reg V1 V2 V3 V4 V5 V6：本命令的含义是以 V1 为因变量，以 V2、V3、V4、V5、V6 为自变量，进行最小二乘回归分析，研究变量之间的因果影响关系。
- logistic V1 V2 V3 V4 V5 V6：本命令的含义是以 V1 为因变量，以 V2、V3、V4、V5、V6 为自变量，进行二元 Logistic 回归分析，研究变量之间的因果影响关系。其中自变量的影响是以优势比（Odds Ratio）的形式输出的。

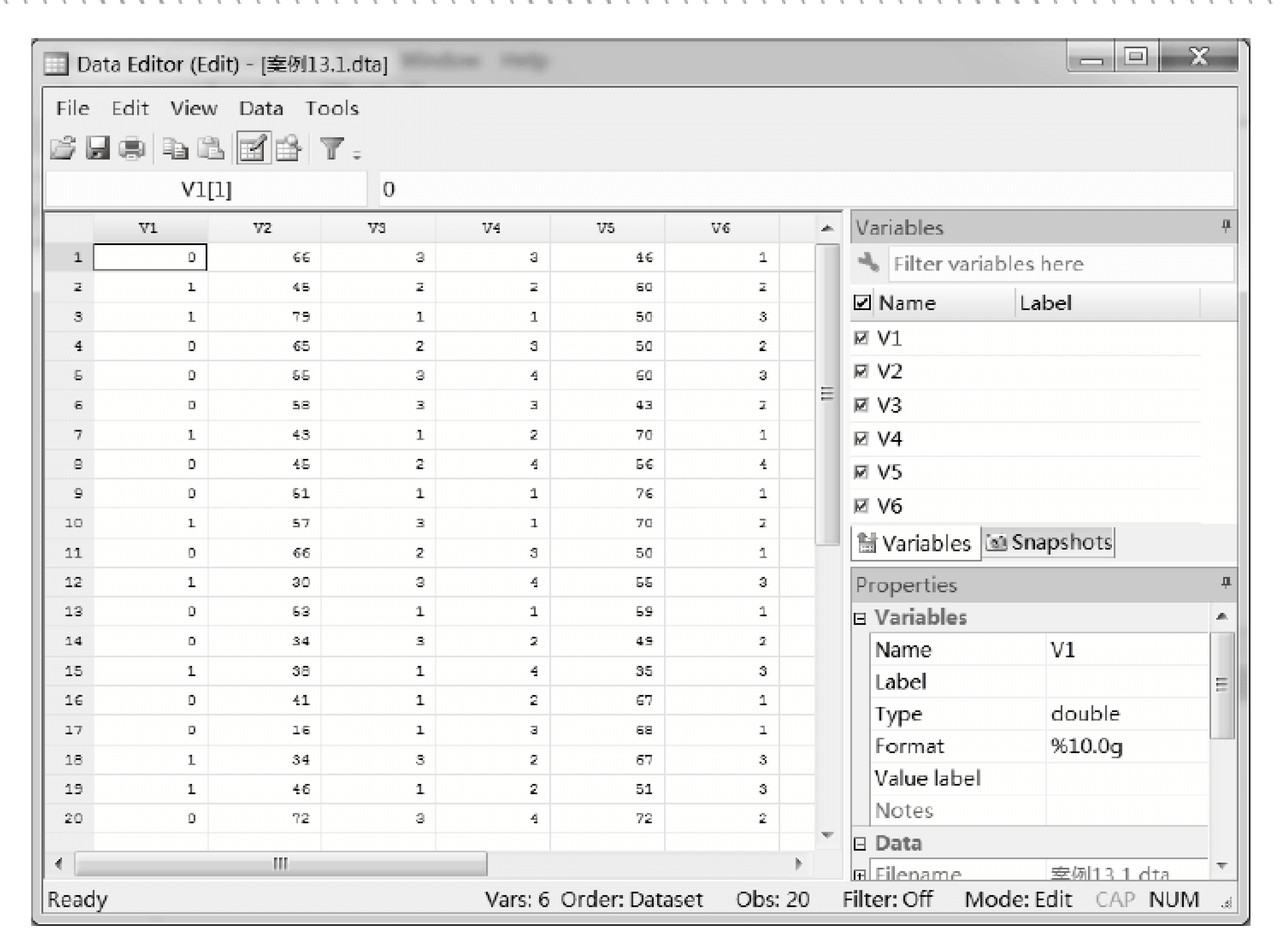


图 13.1 案例 13.1 数据

- logit V1 V2 V3 V4 V5 V6：本命令的含义是以 V1 为因变量，以 V2、V3、V4、V5、V6 为自变量，进行二元 Logistic 回归分析，研究变量之间的因果影响关系。其中自变量的影响是以回归系数的形式输出的。
- estat clas：本命令的含义是计算预测准确的百分比，并提供分类统计和分类表。
- lstat：本命令是上条命令“estat clas”的另一种表达形式。
- predict yhat：本命令旨在估计因变量的拟合值。它创建一个命名为 yhat 的新变量，等于最近一次 Logistic 模型基础上 y=1 的预测概率。
- estat gof：本命令旨在判断模型的拟合效果，或者说模型的解释能力。

03 设置完毕后，按键盘上的回车键，等待输出结果。

13.1.4 结果分析

在 Stata 14.0 主界面的结果窗口可以看到如图 13.2~图 13.9 所示的分析结果。

图 13.2 是对数据进行展示的结果。它的目的是通过对变量所包含的样本数据进行一一展示，以便简单直观地观测出数据的具体特征，为深入分析做好必要准备。

从如图 13.2 所示的分析结果中可以看出，数据的总体质量还是可以的，没有极端异常值，变量间的量纲差距也是可以接受的，可以进入下一步的分析。

图 13.3 是以 V1 为因变量，以 V2、V3、V4、V5、V6 为自变量，进行最小二乘回归分析的结果。

```
. list V1-V6
```

	V1	V2	V3	V4	V5	V6
1.	0	66	3	3	46	1
2.	1	45	2	2	60	2
3.	1	79	1	1	50	3
4.	0	65	2	3	50	2
5.	0	55	3	4	60	3
6.	0	58	3	3	43	2
7.	1	43	1	2	70	1
8.	0	45	2	4	56	4
9.	0	51	1	1	76	1
10.	1	57	3	1	70	2
11.	0	66	2	3	50	1
12.	1	30	3	4	55	3
13.	0	53	1	1	59	1
14.	0	34	3	2	49	2
15.	1	38	1	4	35	3
16.	0	41	1	2	67	1
17.	0	16	1	3	68	1
18.	1	34	3	2	67	3
19.	1	46	1	2	51	3
20.	0	72	3	4	72	2

图 13.2　对数据进行展示

```
. reg V1 V2 V3 V4 V5 V6

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  5,    14) =    1.64
       Model |  1.77379404     5  .354758807           Prob > F      =  0.2135
    Residual |  3.02620596    14  .216157569           R-squared     =  0.3695
-------------+------------------------------           Adj R-squared =  0.1444
       Total |         4.8    19  .252631579           Root MSE      =  .46493

------------------------------------------------------------------------------
          V1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          V2 |  -.0061692    .0072331    -0.85   0.408    -.0216826    .0093441
          V3 |  -.0333053    .1295916    -0.26   0.801    -.3112516    .2446411
          V4 |  -.2071337    .1165346    -1.78   0.097    -.4570756    .0428083
          V5 |  -.0017381    .0108751    -0.16   0.875     -.025063    .0215868
          V6 |   .2997717     .126881     2.36   0.033      .027639    .5719045
       _cons |   .7871698    .9606104     0.82   0.426    -1.273134    2.847474
------------------------------------------------------------------------------
```

图 13.3　最小二乘回归

从上述分析结果中可以看出共有 20 个样本参与了分析，模型的 F 值(5, 14) = 1.64，P 值（Prob > F）= 0.2135，说明模型整体上是不显著的。模型的可决系数（R-squared）为 0.3695，模型修正的可决系数（Adj R-squared）为 0.1444，说明模型的解释能力也是比较差的。

变量 V2 的系数标准误是 0.0072331，t 值为-0.85，P 值为 0.408，系数是不显著的，95%的置信区间为[-0.0216826, 0.0093441]。变量 V3 的系数标准误是 0.1295916，t 值为-0.26，P 值为 0.801，系数是非常不显著的，95%的置信区间为[-0.3112516, 0.2446411]。变量 V4 的系数标准误是 0.1165346，t 值为-1.78，P 值为 0.097，系数的显著性一般，95%的置信区间为[-0.4570756, 0.0428083]。变量 V5 的系数标准误是 0.0108751，t 值为-0.16，P 值为 0.875，系数是非常不显著的，95%的置信区间为[-0.025063, 0.0215868]。变量 V6 的系数标准误是 0.126881，t 值为 2.36，P 值为 0.033，系数是非常显著的，95%的置信区间为[0.027639, 0.5719045]。常数项的系数标准误是 0.9606104，t 值为 0.82，P 值为 0.426，系数也是比较不显著的，95%的置信区间为[-1.273134，2.847474]。

从上述分析结果，我们可以得到最小二乘模型的回归方程是：

V1=-0.0061692*V2-0.0333053*V3 -0.2071337*V4 -0.0017381*V5+0.2997717*V6+0.7871698

从上面的分析可以看出最小二乘线性模型的整体显著性、系数显著性以及模型的整体解释能力都是有较大提升空间的。

图 13.4 是以 V1 为因变量，以 V2、V3、V4、V5、V6 为自变量，进行二元 Logistic 回归分析的结果。其中，自变量的影响是以优势比（Odds Ratio）的形式输出的。

从图 13.4 可以看出 Logistic 模型相对于最小二乘回归模型得到了很大程度的改进。模型的整体显著性 P 值达到了 9%左右（Prob > chi2=0.0934）。伪 R 方达到 35% （Pseudo R2 = 0.3500），解释能力进一步提高。各个变量系数的显著程度也有不同程度的提高，限于篇幅不再赘述。

```
. logistic V1 V2 V3 V4 V5 V6

Logistic regression                               Number of obs   =         20
                                                  LR chi2(5)      =       9.42
                                                  Prob > chi2     =     0.0934
Log likelihood = -8.7492827                       Pseudo R2       =     0.3500

------------------------------------------------------------------------------
          V1 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          V2 |   .9376137   .0546723    -1.10   0.269     .8363544    1.051133
          V3 |   .6501124   .4840099    -0.58   0.563      .151101    2.797111
          V4 |   .2217138    .197692    -1.69   0.091     .0386203    1.272828
          V5 |   .9931318    .068624    -0.10   0.921     .8673414    1.137166
          V6 |   7.819255   8.382102     1.92   0.055     .9565164    63.92022
       _cons |   25.13991   160.7937     0.50   0.614     .0000904     6992105
------------------------------------------------------------------------------
```

图 13.4 二元 Logistic 回归

与一般的回归形式不同，此处自变量的影响是以优势比（Odds Ratio）的形式输出的，它的含义是：在其他自变量保持不变的条件下，被观测自变量每增加 1 个单位时 y=1 的发生比的变化倍数。可以看出，各个变量中只有 V6 变量的增加会引起因变量取 1 值的大于 1 倍的增加。这说明只有 V6 是与因变量呈现正向变化，只有 V6 使得因变量取 1 的概率更大。

图 13.5 是以 V1 为因变量，以 V2、V3、V4、V5、V6 为自变量，进行二元 Logistic 回归分析的结果。其中，自变量的影响是以回归系数的形式输出的。

```
. logit V1 V2 V3 V4 V5 V6

Iteration 0:   log likelihood = -13.460233
Iteration 1:   log likelihood =  -9.046534
Iteration 2:   log likelihood = -8.7562687
Iteration 3:   log likelihood = -8.7492923
Iteration 4:   log likelihood = -8.7492827
Iteration 5:   log likelihood = -8.7492827

Logistic regression                               Number of obs   =         20
                                                  LR chi2(5)      =       9.42
                                                  Prob > chi2     =     0.0934
Log likelihood = -8.7492827                       Pseudo R2       =     0.3500

------------------------------------------------------------------------------
          V1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          V2 |  -.0644172     .05831    -1.10   0.269    -.1787028    .0498683
          V3 |    -.43061   .7445019    -0.58   0.563    -1.889807    1.028587
          V4 |  -1.506368   .8916537    -1.69   0.091    -3.253977    .2412414
          V5 |  -.0068919   .0690986    -0.10   0.921    -.1423226    .1285388
          V6 |   2.056589   1.071982     1.92   0.055    -.0444574    4.157636
       _cons |   3.224457   6.395952     0.50   0.614    -9.311379    15.76029
------------------------------------------------------------------------------
```

图 13.5 自变量的影响以回归系数形式输出

从图 13.5 可以看出该模型与使用 Logistic 命令回归得到的结果是一致的，只是自变量影响输出的形式由优势比换成了回归系数。

最终模型表达式为：

$$LNV1= -0.0644172V2-0.43061V3 -1.506368V4 -0.0068919V5+ 2.056589V6+ 3.224457$$

其中，LNV1、V2、V3、V4、V5、V6 分别表示肾细胞发生癌转移概率的对数值、年龄、肾细胞癌血管内皮生长因子、肾癌细胞核组织学分级、肾细胞癌组织内微血管数和肾细胞癌分期。

综上所述，我们的研究结论是：年龄、肾细胞癌血管内皮生长因子、肾癌细胞核组织学分级、肾细胞癌组织内微血管数与肾细胞癌转移呈反向变化，肾细胞癌分期与肾细胞癌转移呈正向变化，但这些变化并不是特别显著。

图 13.6 是计算预测准确的百分比，并提供分类统计和分类表的结果。

```
. estat clas

Logistic model for V1

              ──────── True ────────
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |         6             2  |          8
     -     |         2            10  |         12
-----------+--------------------------+-----------
   Total   |         8            12  |         20

Classified + if predicted Pr(D) >= .5
True D defined as V1 != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   75.00%
Specificity                     Pr( -|~D)   83.33%
Positive predictive value       Pr( D| +)   75.00%
Negative predictive value       Pr(~D| -)   83.33%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   16.67%
False - rate for true D         Pr( -| D)   25.00%
False + rate for classified +   Pr(~D| +)   25.00%
False - rate for classified -   Pr( D| -)   16.67%
--------------------------------------------------
Correctly classified                        80.00%
--------------------------------------------------
```

图 13.6　计算预测准确的百分比

从图 13.6 可以看出很多信息。按照系统默认设置，系统使用 0.5 作为分割点。分类中的 D、-D、“+”和“-”分别表示的含义如下。

- D：表示一个观测样本所关注的事件的确发生了，也就是说 y 的值取到了 1，在本例中，也就是说肾细胞确实发生了癌转移。
- -D：表示一个观测样本所关注的事件的确没有发生，也就是说 y 的值取到了 0，在本例中，也就是说肾细胞确实没有发生癌转移。
- +：表示模型预测的概率值大于分割点，本例中，也就是说模型预测的肾细胞发生癌转移的概率为 0.5 或者更多。
- -：表示模型预测的概率值小于分割点，本例中，也就是说模型预测的肾细胞发生癌转移的概率低于 0.5。

所以，按照模型预测肾细胞发生癌转移的概率至少在 0.5 以上的标准，有 6 次是肾细胞确实发生了癌转移而且模型预测的概率值大于分割点，有 10 次是肾细胞确实没有发生癌转移而且模型预测的概率值小于分割点，所以，一共有 16 个样本的预测是正确的，预测正确率占全部样本的百分之八十（80%）。有 2 次是肾细胞确实发生了癌转移但模型预测的概率值小于分割点，有 2 次是肾细胞确实没有发生癌转移但模型预测的概率值大于分割点，一共有 4 个样本的预测是错误的，预测错误率占全部样本的百分之二十（20%）。

图 13.7 是上条命令“estat clas”的另一种表达形式的结果。该结果与图 13.6 的结果一致。

```
. lstat

Logistic model for V1

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |         6             2  |          8
     -     |         2            10  |         12
-----------+--------------------------+-----------
   Total   |         8            12  |         20

Classified + if predicted Pr(D) >= .5
True D defined as V1 != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   75.00%
Specificity                     Pr( -|~D)   83.33%
Positive predictive value       Pr( D| +)   75.00%
Negative predictive value       Pr(~D| -)   83.33%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   16.67%
False - rate for true D         Pr( -| D)   25.00%
False + rate for classified +   Pr(~D| +)   25.00%
False - rate for classified -   Pr( D| -)   16.67%
--------------------------------------------------
Correctly classified                        80.00%
--------------------------------------------------
```

图 13.7 分析结果图

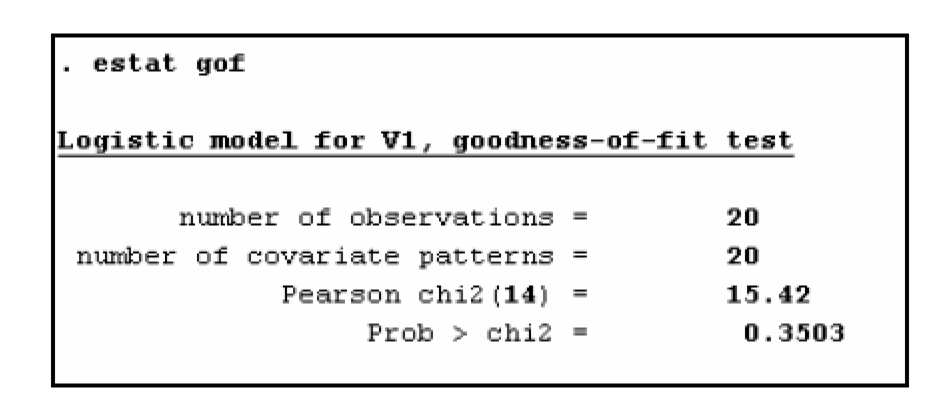
图 13.8 是对因变量的拟合值的预测。选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 13.8 所示的 yhat 数据。

	V1	V2	V3	V4	V5	V6	yhat
1	0	66	3	3	46	1	.0060691
2	1	45	2	2	60	2	.5377755
3	1	79	1	1	50	3	.883262
4	0	65	2	3	50	2	.0708041
5	0	55	3	4	60	3	.1324419
6	0	58	3	3	43	2	.0754491
7	1	43	1	2	70	1	.195498
8	0	45	2	4	56	4	.7823478
9	0	51	1	1	76	1	.3858016
10	1	57	3	1	70	2	.5951321
11	0	66	2	3	50	1	.0090544
12	1	30	3	4	55	3	.4416058
13	0	53	1	1	59	1	.3830394
14	0	34	3	2	49	2	.6236761
15	1	38	1	4	35	3	.5619481
16	0	41	1	2	67	1	.2200864
17	0	16	1	3	68	1	.2372182
18	1	34	3	2	67	3	.9196585
19	1	46	1	2	51	3	.9331553
20	0	72	3	4	72	2	.0059765

图 13.8 对变量拟合值的预测

二元 Logistic 的因变量拟合值预测结果表示的含义是 y=1 的概率，本例所表示的含义是肾细胞发生癌转移的概率。

图 13.9 是对 Logistic 模型拟合效果的分析结果。



```
. estat gof

Logistic model for V1, goodness-of-fit test

       number of observations =        20
 number of covariate patterns =        20
             Pearson chi2(14) =        15.42
                  Prob > chi2 =         0.3503
```

图 13.9 对 Logistic 模型拟合效果的分析结果

可以看到 Prob > chi2 =0.3503，说明模型的解释能力还是差强人意的，但比最小二乘线性回归模型要好出很多。

13.1.5 案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

1. 延伸 1：设定模型预测概率的具体值

我们在上述分析过程和结果分析中都用的是系统默认设置的 0.5 概率对模型估计有效性进行的评价。事实上，我们完全可以自由设定需要的概率水平对模型做出评价。例如，我们要求预测概率达到 80%，那么操作命令就是：

```
estat clas,cutoff(0.8)r
```

在命令窗口输入命令并按回车键进行确认，结果如图 13.10 所示。

```
. estat clas,cutoff(0.8)

Logistic model for V1

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |         3             0  |          3
     -     |         5            12  |         17
-----------+--------------------------+-----------
   Total   |         8            12  |         20

Classified + if predicted Pr(D) >= .8
True D defined as V1 != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   37.50%
Specificity                     Pr( -|~D)  100.00%
Positive predictive value       Pr( D| +)  100.00%
Negative predictive value       Pr(~D| -)   70.59%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    0.00%
False - rate for true D         Pr( -| D)   62.50%
False + rate for classified +   Pr(~D| +)    0.00%
False - rate for classified -   Pr( D| -)   29.41%
--------------------------------------------------
Correctly classified                        75.00%
--------------------------------------------------
```

图 13.10 设定模型预测概率的具体值

从上面的分析结果中可以看出在设置概率为 0.8 的时候，模型的预测正确性降到了 75%。读者可以自行设定其他的概率水平继续进行深入研究。

2. 延伸 2：使用 probit 模型对二分类因变量进行拟合

以本节中介绍的实例进行说明，那么操作命令如下。

（1）probit V1 V2 V3 V4 V5 V6

本命令的含义是以 V1 为因变量，以 V2、V3、V4、V5、V6 为自变量，进行 probit 回归分析，研究变量之间的因果影响关系。

（2）mfx

本命令旨在计算在样本均值处的边际效应。

（3）estat clas

本命令的含义是计算预测准确的百分比，并提供分类统计和分类表。

（4）predict yhat

本命令旨在估计因变量的拟合值。它创建一个命名为 yhat 的新变量，等于最近一次 Probit 模型基础上 y=1 的预测概率。

在命令窗口输入命令并按回车键进行确认，结果如图 13.11~图 13.14 所示。

图 13.11 是以 V1 为因变量，以 V2、V3、V4、V5、V6 为自变量，进行 Probit 回归分析的结果。

```
. probit V1 V2 V3 V4 V5 V6

Iteration 0:   log likelihood = -13.460233
Iteration 1:   log likelihood = -8.8919351
Iteration 2:   log likelihood = -8.6758218
Iteration 3:   log likelihood = -8.6723658
Iteration 4:   log likelihood = -8.6723655

Probit regression                                 Number of obs   =         20
                                                  LR chi2(5)      =       9.58
                                                  Prob > chi2     =     0.0882
Log likelihood = -8.6723655                       Pseudo R2       =     0.3557

------------------------------------------------------------------------------
          V1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          V2 |  -.0387215   .0346081    -1.12   0.263    -.1065521    .0291092
          V3 |  -.2637185   .4397551    -0.60   0.549    -1.125623    .5981057
          V4 |  -.9287975   .5247439    -1.77   0.077    -1.957277    .0996817
          V5 |  -.0049234   .0403986    -0.12   0.903    -.0841032    .0742565
          V6 |   1.227209   .5931853     2.07   0.039     .0645876    2.389831
       _cons |   2.064971   3.774138     0.55   0.584    -5.332204    9.462146
------------------------------------------------------------------------------
```

图 13.11　Probit 回归

从上面的分析结果中可以看出，Probit 模型与 Logistic 模型所得的结果相差不大，模型整体的显著程度和解释能力都相比最小二乘回归分析有所提高。

图 13.12 是在样本均值处的边际效应结果。

```
. mfx

Marginal effects after probit
      y  = Pr(V1) (predict)
         =  .30025942
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
      V2 |  -.0134684      .01067   -1.26   0.207  -.034375  .007438     49.7
      V3 |  -.0917288      .14875   -0.62   0.537  -.383269  .199811        2
      V4 |  -.3230623      .16421   -1.97   0.049    -.6449 -.001224     2.55
      V5 |  -.0017125      .01408   -0.12   0.903  -.029307  .025882     57.7
      V6 |   .4268584      .18498    2.31   0.021   .064298  .789419     2.05
------------------------------------------------------------------------------
```

图 13.12　在样本均值处的边际效应结果

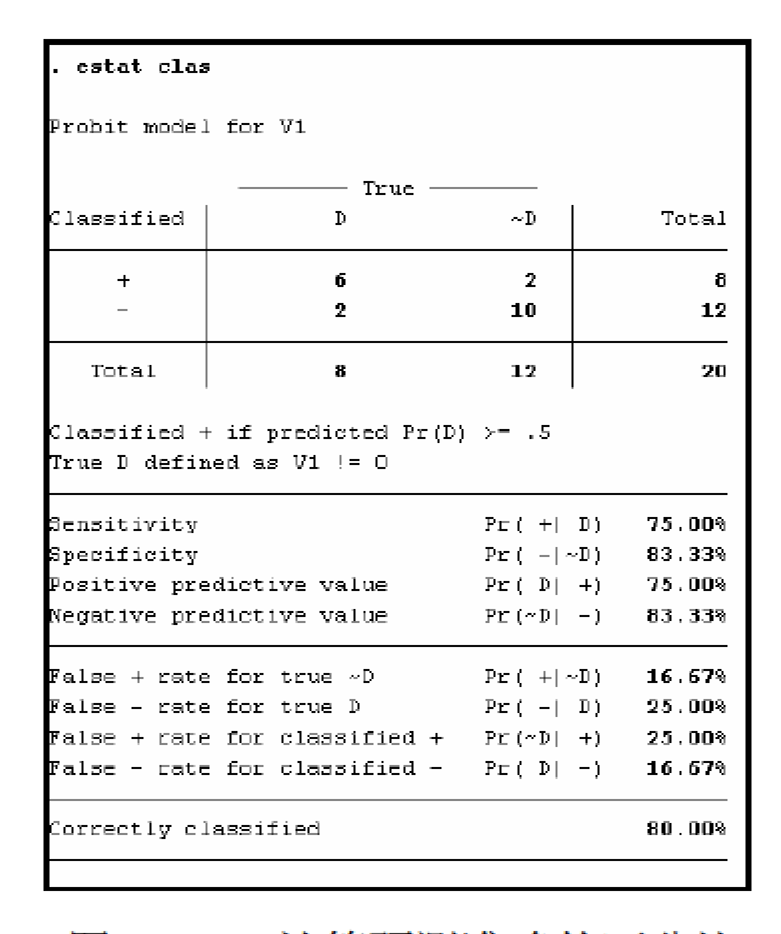
从图 13.12 可以看出，Probit 模型在样本均值处的边际效应与最小二乘回归分析相差不大。

图 13.13 是计算预测准确的百分比，并提供分类统计和分类表的结果。

从图 13.13 可以看出预测正确率占全部样本的百分之八十（80%），这与 Logistic 模型得到的结论是相同的。

图 13.14 是对因变量的拟合值的预测。选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 13.14 所示的 yhat 数据。

与 Logistic 模型相同，Probit 模型的因变量拟合值预测结果表示的含义也是 y=1 的概率，本例所表示的含义同样是肾细胞发生癌转移的概率。



```
. estat clas

Probit model for V1

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |         6             2  |          8
     -     |         2            10  |         12
-----------+--------------------------+-----------
   Total   |         8            12  |         20

Classified + if predicted Pr(D) >= .5
True D defined as V1 != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   75.00%
Specificity                     Pr( -|~D)   83.33%
Positive predictive value       Pr( D| +)   75.00%
Negative predictive value       Pr(~D| -)   83.33%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   16.67%
False - rate for true D         Pr( -| D)   25.00%
False + rate for classified +   Pr(~D| +)   25.00%
False - rate for classified -   Pr( D| -)   16.67%
--------------------------------------------------
Correctly classified                        80.00%
--------------------------------------------------
```

图 13.13　计算预测准确的百分比

	V1	V2	V3	V4	V5	V6	yhat
1	0	66	3	3	46	1	.0010794
2	1	45	2	2	60	2	.5384346
3	1	79	1	1	50	3	.8941529
4	0	65	2	3	50	2	.0596756
5	0	55	3	4	60	3	.1180427
6	0	58	3	3	43	2	.0647967
7	1	43	1	2	70	1	.2007934
8	0	45	2	4	56	4	.7620795
9	0	51	1	1	76	1	.4015624
10	1	57	3	1	70	2	.5978085
11	0	66	2	3	50	1	.0023756
12	1	30	3	4	55	3	.4238026
13	0	53	1	1	59	1	.403983
14	0	34	3	2	49	2	.6228085
15	1	38	1	4	35	3	.5493255
16	0	41	1	2	67	1	.2276592
17	0	16	1	3	68	1	.2381512
18	1	34	3	2	67	3	.9266733
19	1	46	1	2	51	3	.9444207
20	0	72	3	4	72	2	.0008759

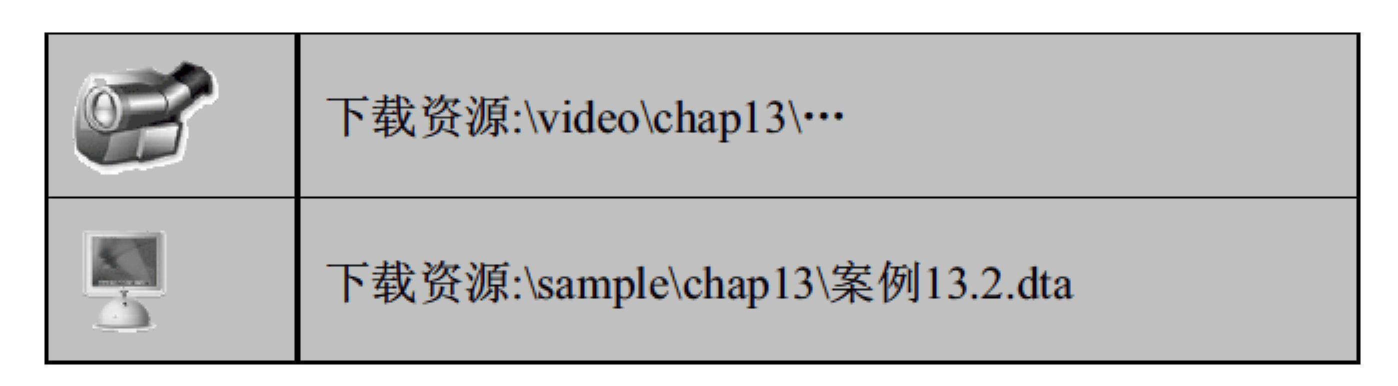
图 13.14　因变量的拟合值预测

13.2　实例二——多元Logistic回归分析

13.2.1　多元 Logistic 回归分析的功能与意义

我们经常会遇到因变量有多个取值而且无大小顺序的情况，例如职业、婚姻情况等，这时一般的线性回归分析无法准确地刻画变量之间的因果关系，需要用其他的回归分析方法来进行拟合模型。Stata 的多项分类 Logistic 回归便是一种简便的处理该类因变量问题的分析方法。

13.2.2　相关数据来源



下载资源:\video\chap13\…

下载资源:\sample\chap13\案例13.2.dta

【例 13.2】表 13.2 给出了对山东省某中学 20 名视力低下学生视力监测的结果数据。试用多项分类 Logistic 回归方法分析视力低下程度（由轻到重共 3 级）与年龄、性别（1 代表男性，2 代表女性）之间的关系。

表 13.2　山东省某中学 20 名学生视力监测结果数据

编号	视力低下程度	性别	年龄
1	1	1	15
2	1	1	15
3	2	1	14
4	2	2	16
5	3	2	16
6	3	2	17
7	2	2	17

（续表）

编号	视力低下程度	性别	年龄
8	2	1	18
9	1	1	14
10	3	2	18
11	1	1	17
12	1	2	17
13	1	1	15
14	2	1	18
15	1	2	15
16	1	2	15
17	3	2	17
18	1	1	15
19	1	1	15
20	2	2	16

13.2.3 Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有 3 个变量，分别是视力低下程度、性别和年龄。我们把视力低下程度变量设定为 V1，把性别变量设定为 V2，把年龄变量设定为 V3，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作我们在第 1 章中已有详细讲述。录入完成后数据如图 13.15 所示。

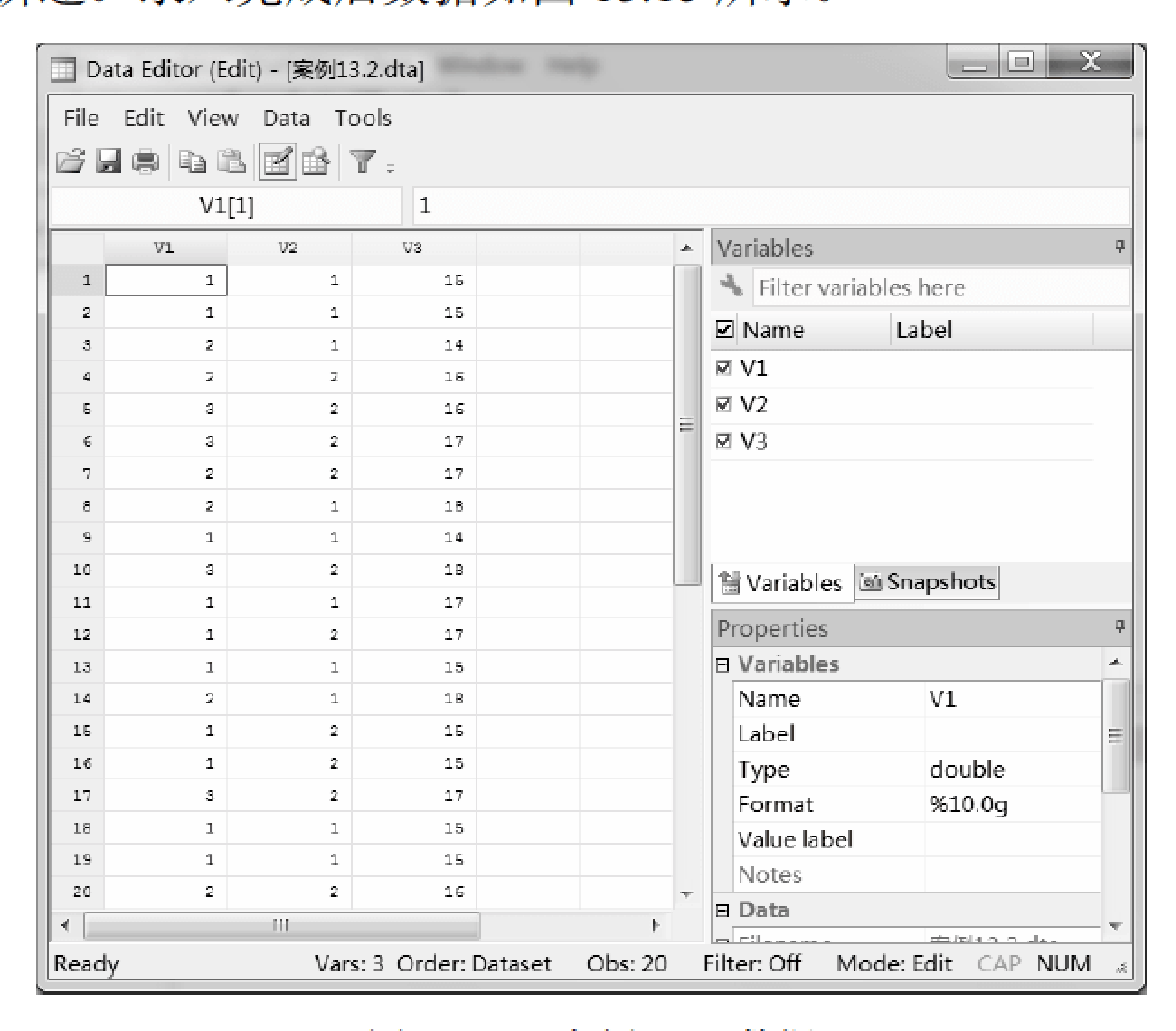


图 13.15 案例 13.2 数据

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的"Command"文本框中输入操作命令并按键盘上的回车键进行确认。本例中提到的各步要求对应的命令分别如下。

- list V1-V3：本命令的含义是对 3 个变量所包含的样本数据进行一一展示，以便简单直观地观测出数据的具体特征，为深入分析做好必要准备。
- reg V1 V2 V3：本命令的含义是以 V1 为因变量，以 V2、V3 为自变量，进行最小二乘回归分析，研究变量之间的因果影响关系。
- mlogit V1 V2 V3,base(1)：本命令的含义是以 V1 为因变量，以 V2、V3 为自变量，并设定第 1 组为参照组（视力低下程度为 1），进行多元 Logistic 回归分析，研究变量之间的因果影响关系。其中自变量的影响是以回归系数的形式输出的。
- mlogit V1 V2 V3,base(1) rrr：本命令的含义是以 V1 为因变量，以 V2、V3 为自变量，并设定第 1 组为参照组（视力低下程度为 1），进行多元 Logistic 回归分析，研究变量之间的因果影响关系。其中，自变量的影响是以相对风险比率的形式输出的。

13.2.4 结果分析

在 Stata 14.0 主界面的结果窗口可以看到如图 13.16~图 13.19 所示的分析结果。

图 13.16 是对数据进行展示的结果。它的目的是通过对变量所包含的样本数据进行一一展示，以便简单直观地观测出数据的具体特征，为深入分析做好必要准备。

```
. list V1-V3

     +--------------+
     | V1   V2   V3 |
     |--------------|
  1. |  1    1   15 |
  2. |  1    1   15 |
  3. |  2    1   14 |
  4. |  2    2   16 |
  5. |  3    2   16 |
     |--------------|
  6. |  3    2   17 |
  7. |  2    2   17 |
  8. |  2    1   18 |
  9. |  1    1   14 |
 10. |  3    2   18 |
     |--------------|
 11. |  1    1   17 |
 12. |  1    2   17 |
 13. |  1    1   15 |
 14. |  2    1   18 |
 15. |  1    2   15 |
     |--------------|
 16. |  1    2   15 |
 17. |  3    2   17 |
 18. |  1    1   15 |
 19. |  1    1   15 |
 20. |  2    2   16 |
     +--------------+
```

图 13.16　对数据进行展示

在如图 13.16 所示的分析结果中可以看出，数据的总体质量还是可以的，没有极端异常值，变量间的量纲差距也是可以接受的，可以进入下一步的分析。

图 13.17 是以 V1 为因变量，以 V2、V3 为自变量，进行最小二乘回归分析的结果。

```
. reg V1 V2 V3

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  2,    17) =    6.56
       Model |     5.3125     2     2.65625           Prob > F      =  0.0078
    Residual |     6.8875    17  .405147059           R-squared     =  0.4355
-------------+------------------------------           Adj R-squared =  0.3690
       Total |       12.2    19  .642105263           Root MSE      =  .63651

------------------------------------------------------------------------------
          V1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          V2 |   .5833333   .3000545     1.94   0.069    -.0497262    1.216393
          V3 |   .2708333   .1186069     2.28   0.036     .0205946    .5210721
       _cons |  -3.508333   1.812165    -1.94   0.070    -7.331667    .3150006
------------------------------------------------------------------------------
```

图 13.17 最小二乘回归分析

从上述分析结果中可以看出共有 20 个样本参与了分析，模型的 F 值(2, 17) = 6.56，P 值（Prob > F） = 0.0078，说明模型整体上是比较显著的。模型的可决系数（R-squared）= 0.4355，模型修正的可决系数（Adj R-squared）= 0.3690，说明模型的解释能力差强人意。

变量 V2 的系数标准误是 0.3000545，t 值为 1.94，P 值为 0.069，系数显著性是勉强过得去的，95%的置信区间为[-0.0497262, 1.216393]。变量 V3 的系数标准误是 0.1186069，t 值为 2.28，P 值为 0.036，系数是比较显著的，95%的置信区间为[0.0205946, 0.5210721]。常数项的系数标准误是 1.812165，t 值为-1.94，P 值为 0.070，系数显著性是勉强过得去的，95%的置信区间为[-7.331667,0.3150006]。

从上述分析结果可以得到最小二乘模型的回归方程是：

V1=0.583333 *V2 +0.2708333 *V3-3.508333

从上面的分析可以看出最小二乘线性模型的整体显著性和系数显著性以及模型的整体解释能力都是勉强过得去的。

图 13.18 是以 V1 为因变量，以 V2、V3 为自变量，并设定第 1 组为参照组（视力低下程度为 1），进行多元 Logistic 回归分析的结果。其中，自变量的影响是以回归系数的形式输出的。

```
. mlogit  V1 V2 V3,base(1)

Iteration 0:   log likelihood =  -20.59306
Iteration 1:   log likelihood = -15.340101
Iteration 2:   log likelihood =  -14.03923
Iteration 3:   log likelihood = -13.734306
Iteration 4:   log likelihood =  -13.69158
Iteration 5:   log likelihood = -13.681016
Iteration 6:   log likelihood = -13.679506
Iteration 7:   log likelihood = -13.679011
Iteration 8:   log likelihood = -13.678908
Iteration 9:   log likelihood = -13.678885
Iteration 10:  log likelihood = -13.678879
Iteration 11:  log likelihood = -13.678878

Multinomial logistic regression                   Number of obs   =         20
                                                  LR chi2(4)      =      13.83
                                                  Prob > chi2     =     0.0079
Log likelihood = -13.678878                       Pseudo R2       =     0.3358

------------------------------------------------------------------------------
          V1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |  (base outcome)
-------------+----------------------------------------------------------------
2            |
          V2 |    .732262   1.183462     0.62   0.536    -1.587281    3.051805
          V3 |   .8356566   .4982461     1.68   0.094    -.1408878    1.812201
       _cons |  -14.82979   8.211396    -1.81   0.071    -30.92383    1.264249
-------------+----------------------------------------------------------------
3            |
          V2 |   18.39871   1982.115     0.01   0.993    -3866.474    3903.272
          V3 |   2.112522   1.181372     1.79   0.074    -.2029232    4.427968
       _cons |  -71.13788   3964.291    -0.02   0.986    -7841.005    7698.729
------------------------------------------------------------------------------
```

图 13.18 多元 Logistic 回归分析

从图 13.18 可以看出 Logistic 模型与最小二乘回归估计效果相差不大。模型的整体显著性 P 值达到了 0.0079 左右（Prob > chi2=0.0079）。伪 R 方达到 33.58% （Pseudo R2 = 0.3358），解释能力进一步提高。

从图 13.18 中可以看到 V2 和 V3 系数在第 2 组和第 3 组都是大于 0 的，这意味着 V2 和 V3 两个变量的值越大就越容易被分到 2、3 组，这表示性别为女，年龄越大，越容易被分到中度视力低下、重度视力低下组。

最终模型方程为：

- G1=0，因为轻度是因变量中的参考组，其所有系数均为 0。
- G2=LOG[P(低下中度)/P(低下轻度)]= -14.82979+0.8356566*年龄+0.732262*性别 1。
- G3= LOG[P(低下重度)/P(低下轻度)]= -71.13788+2. 112522*年龄+18.39871*性别 1。

图 13.19 是以 V1 为因变量，以 V2、V3 为自变量，进行多元 Logistic 回归分析的结果。其中，自变量的影响是以相对风险比率的形式输出的。

```
. mlogit  V1 V2 V3,base(1) rrr

Iteration 0:   log likelihood =  -20.59306
Iteration 1:   log likelihood = -15.348101
Iteration 2:   log likelihood =  -14.03923
Iteration 3:   log likelihood = -13.734306
Iteration 4:   log likelihood =  -13.69158
Iteration 5:   log likelihood = -13.681816
Iteration 6:   log likelihood = -13.679506
Iteration 7:   log likelihood = -13.679011
Iteration 8:   log likelihood = -13.678908
Iteration 9:   log likelihood = -13.678885
Iteration 10:  log likelihood = -13.678879
Iteration 11:  log likelihood = -13.678878

Multinomial logistic regression                   Number of obs   =         20
                                                  LR chi2(4)      =      13.83
                                                  Prob > chi2     =     0.0079
Log likelihood = -13.678878                       Pseudo R2       =     0.3358

------------------------------------------------------------------------------
          V1 |        RRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |  (base outcome)
-------------+----------------------------------------------------------------
2            |
          V2 |    2.07978   2.461341     0.62   0.536     .2044808     21.1535
          V3 |   2.306328   1.149119     1.68   0.094     .8685868    6.123911
       _cons |   3.63e-07   2.98e-06    -1.81   0.071     3.71e-14    3.540432
-------------+----------------------------------------------------------------
3            |
          V2 |   9.78e+07   1.94e+11     0.01   0.993            0           .
          V3 |   8.269073   9.768848     1.79   0.074     .8163409    83.76105
       _cons |   1.27e-31   5.05e-28    -0.02   0.986            0           .
------------------------------------------------------------------------------
```

图 13.19　自变量的影响以相对风险比率的形式输出

与二元 Logistic 中的优势比（Odds Ratio）的概念类似，相对风险比率的含义是：在其他自变量保持不变的条件下，被观测自变量每增加 1 个单位时 y=1 的发生比的变化倍数。可以看出，当 V2 增加或者说性别为女生时，它会有相当大的概率被分到第 3 组，即重度视力低下，当年龄偏大时，它也有较大的概率被分到第 3 组，即重度视力低下。

13.2.5　案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

延伸：根据模型预测每个观测样本视力低下程度的可能性

以本节中介绍的实例进行说明，那么操作命令就是：

```
predict eye1 eye2 eye3
```

图 13.20 是根据模型预测每个观测样本视力低下程度的可能性的结果。选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 13.20 所示的 eye1~eye3 数据。

	V1	V2	V3	eye1	eye2	eye3
1	1	1	15	.8267376	.1732624	5.96e-10
2	1	1	15	.8267376	.1732624	5.96e-10
3	2	1	14	.9167004	.0832996	7.99e-11
4	2	2	16	.3864073	.3884369	.2251557
5	3	2	16	.3864073	.3884369	.2251557
6	3	2	17	.1228992	.2849346	.5921661
7	2	2	17	.1228992	.2849346	.5921661
8	2	1	18	.2800346	.7199653	1.14e-07
9	1	1	14	.9167004	.0832996	7.99e-11
10	3	2	18	.0216497	.1157628	.8625875
11	1	1	17	.4728684	.5271316	2.33e-08
12	1	2	17	.1228992	.2849346	.5921661
13	1	1	15	.8267376	.1732624	5.96e-10
14	2	1	18	.2800346	.7199653	1.14e-07
15	1	2	15	.6638637	.2893564	.0467799
16	1	2	15	.6638637	.2893564	.0467799
17	3	2	17	.1228992	.2849346	.5921661
18	1	1	15	.8267376	.1732624	5.96e-10
19	1	1	15	.8267376	.1732624	5.96e-10
20	2	2	16	.3864073	.3884369	.2251557

图 13.20　根据模型预测样本视力低下程度

如图 13.20 所示，第 1 个观测样本为男性，15 岁，他有 80%以上的概率进入第 1 组，即轻度视力低下，有极小的甚至可以忽略不计的概率被分到第 3 组，即重度视力低下。其他的观测样本，读者可以按照类似的方法逐一进行分析，可以看出，我们的模型构建的不错，模型的预测能力也是比较优秀的。

13.3　实例三——有序Logistic回归分析

13.3.1　有序 Logistic 回归分析的功能与意义

在有些分析研究中，因变量虽然离散但存在着一定的排序，例如消费者对服务行业满意度的评价（很满意、基本满意、不满意、很不满意），又例如消费者对某种品牌产品的忠诚度的衡量（很喜欢、比较喜欢、不喜欢、很不喜欢）。在上述情况下，使用普通最小二乘回归分析以及二元或多元 Logistic 回归分析都不能获得比较好的效果，这时就需要用到我们本节介绍的有序 Logistic 回归分析。

13.3.2　相关数据来源

	下载资源:\video\chap13\…
	下载资源:\sample\chap13\案例13.3.dta

【**例 13.3**】为了获得消费者的满意度情况，某公司对 120 位随机抽取的消费者进行了调

查，其中回收有效样本 114 个，相关信息如表 13.3 所示。试用有序 Logistic 回归方法分析消费者满意程度（1 表示很满意，2 表示基本满意，3 表示不满意）与性别（1 代表男性，2 代表女性）、学历（1 表示大学专科及以下，2 表示大学本科，3 表示研究生及以上）之间的关系。

表 13.3　某公司调查的 114 位消费者信息情况数据

编号	消费者满意程度	性别	学历
1	1	1	1
2	1	1	1
3	2	1	1
4	2	2	1
5	3	2	2
6	3	2	2
…	…	…	…
109	2	1	2
110	3	2	3
111	1	1	1
112	2	1	2
113	3	2	3
114	1	1	2

13.3.3　Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有 3 个变量，分别是消费者满意程度、性别和学历。我们把消费者满意程度变量设定为 V1，把性别变量设定为 V2，把学历变量设定为 V3，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作我们在第 1 章中已有详细讲述。录入完成后数据如图 13.21 所示。

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入操作命令并按键盘上的回车键进行确认。本例中提到的各步要求对应的命令分别如下。

- list V1-V3：本命令的含义是对 3 个变量所包含的样本数据进行一一展示，以便简单直观地观测出数据的具体特征，为深入分析做好必要准备。
- reg V1 V2 V3：本命令的含义是以 V1 为因变量，以 V2、V3 为自变量，进行最小二乘回归分析，研究变量之间的因果影响关系。
- ologit　V1 V2 V3：本命令的含义是以 V1 为因变量，以 V2、V3 为自变量，进行有序 Logistic 回归分析，研究变量之间的因果影响关系。
- predict satisfy1 satisfy2 satisfy3：本命令的含义是根据模型预测每个观测样本满意程度的可能性的结果。

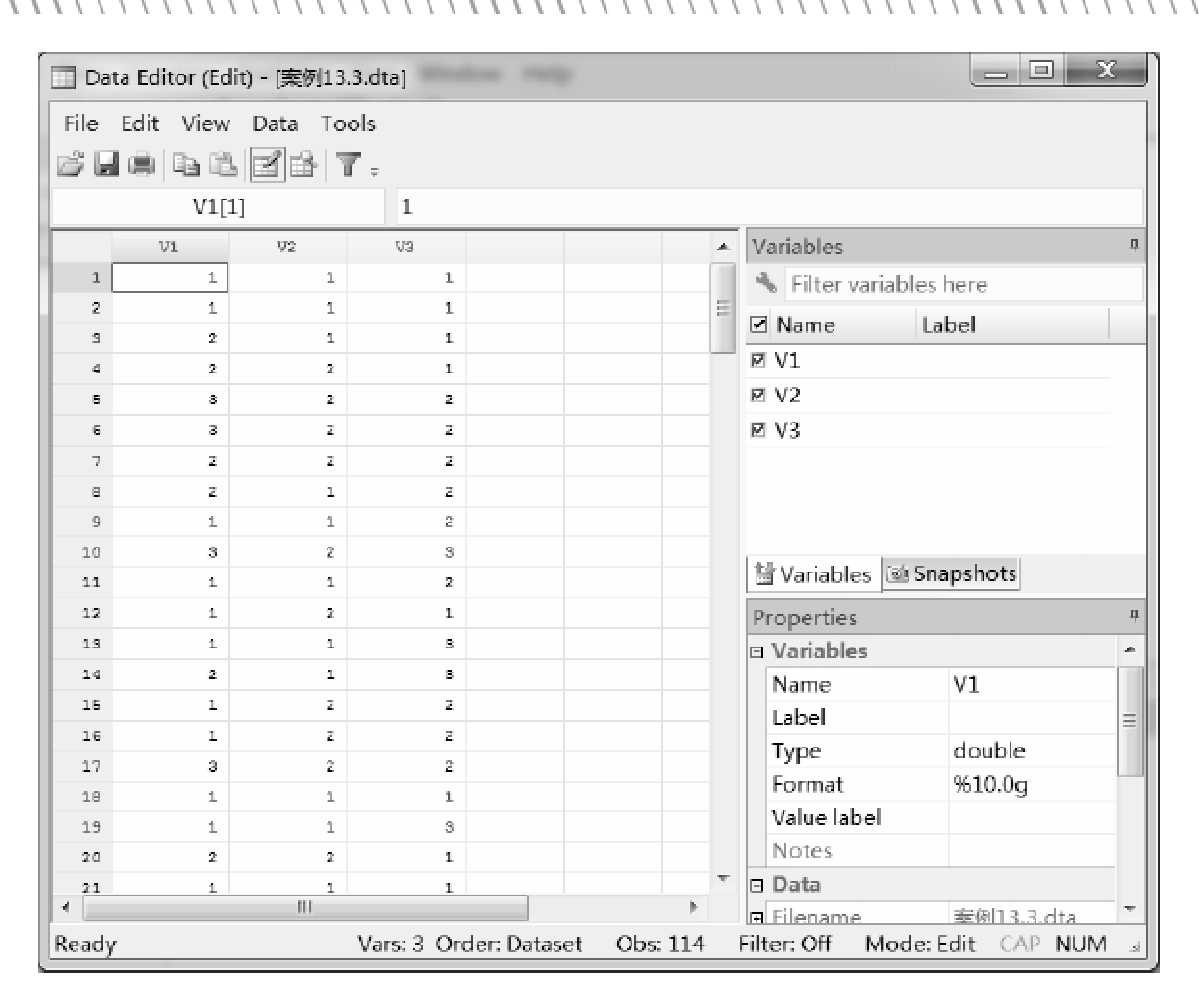
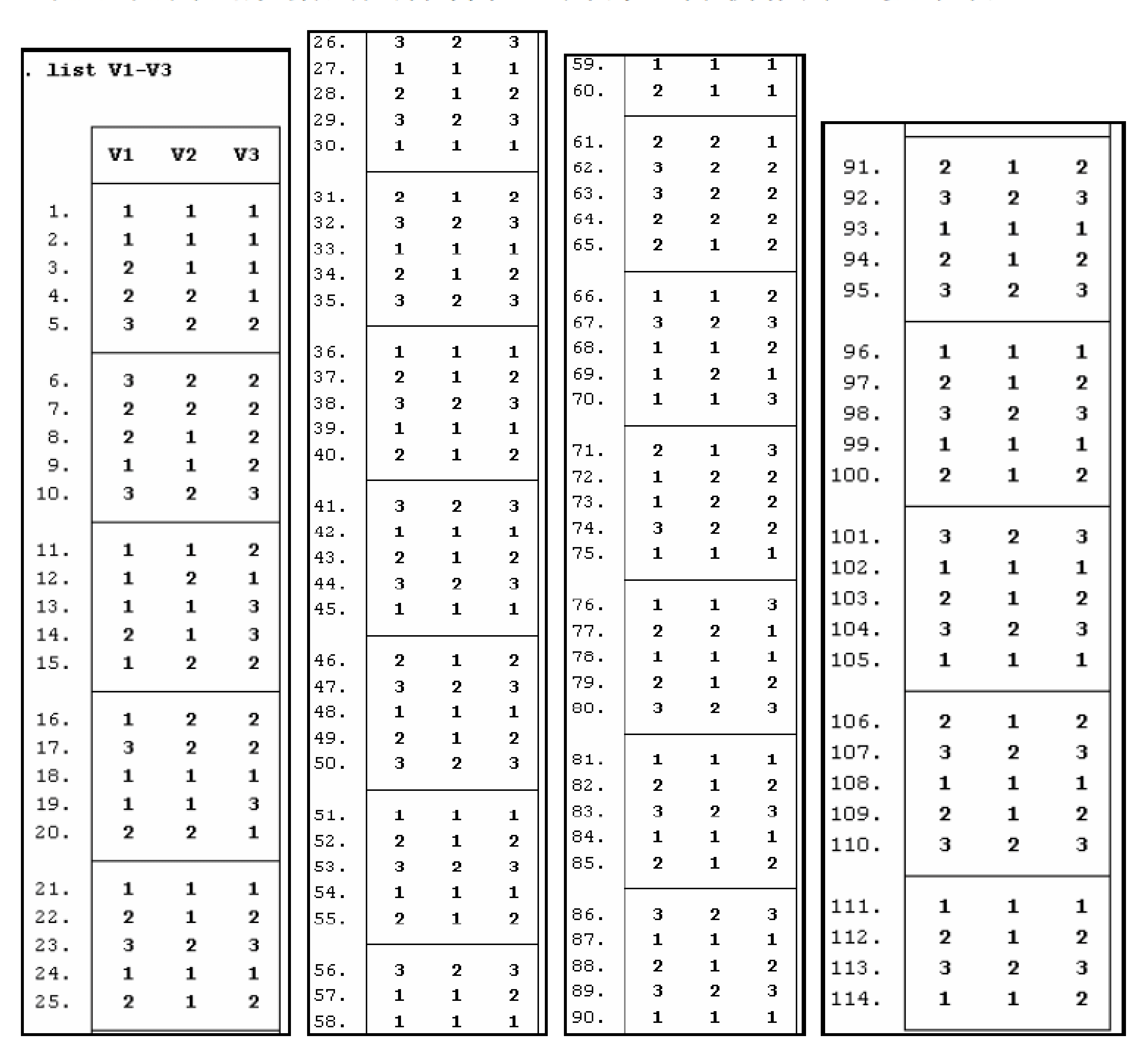


图 13.21 案例 13.3 数据

13.3.4 结果分析

在 Stata 14.0 主界面的结果窗口可以看到如图 13.22~图 13.25 所示的分析结果。

图 13.22 是对数据进行展示的结果。它的目的是通过对变量所包含的样本数据进行一一展示，以便简单直观地观测出数据的具体特征，为深入分析做好必要准备。



. list V1-V3

	V1	V2	V3
1.	1	1	1
2.	1	1	1
3.	2	1	1
4.	2	2	1
5.	3	2	2
6.	3	2	2
7.	2	2	2
8.	2	1	2
9.	1	1	2
10.	3	2	3
11.	1	1	2
12.	1	2	1
13.	1	1	3
14.	2	1	3
15.	1	2	2
16.	1	2	2
17.	3	2	2
18.	1	1	1
19.	1	1	3
20.	2	2	1
21.	1	1	1
22.	2	1	2
23.	3	2	3
24.	1	1	1
25.	2	1	2
26.	3	2	3
27.	1	1	1
28.	2	1	2
29.	3	2	3
30.	1	1	1
31.	2	1	2
32.	3	2	3
33.	1	1	1
34.	2	1	2
35.	3	2	3
36.	1	1	1
37.	2	1	2
38.	3	2	3
39.	1	1	1
40.	2	1	2
41.	3	2	3
42.	1	1	1
43.	2	1	2
44.	3	2	3
45.	1	1	1
46.	2	1	2
47.	3	2	3
48.	1	1	1
49.	2	1	2
50.	3	2	3
51.	1	1	1
52.	2	1	2
53.	3	2	3
54.	1	1	1
55.	2	1	2
56.	3	2	3
57.	1	1	2
58.	1	1	1
59.	1	1	1
60.	2	1	1
61.	2	2	1
62.	3	2	2
63.	3	2	2
64.	2	2	2
65.	2	1	2
66.	1	1	2
67.	3	2	3
68.	1	1	2
69.	1	2	1
70.	1	1	3
71.	2	1	3
72.	1	2	2
73.	1	2	2
74.	3	2	2
75.	1	1	1
76.	1	1	3
77.	2	2	1
78.	1	1	1
79.	2	1	2
80.	3	2	3
81.	1	1	1
82.	2	1	2
83.	3	2	3
84.	1	1	1
85.	2	1	2
86.	3	2	3
87.	1	1	1
88.	2	1	2
89.	3	2	3
90.	1	1	1
91.	2	1	2
92.	3	2	3
93.	1	1	1
94.	2	1	2
95.	3	2	3
96.	1	1	1
97.	2	1	2
98.	3	2	3
99.	1	1	1
100.	2	1	2
101.	3	2	3
102.	1	1	1
103.	2	1	2
104.	3	2	3
105.	1	1	1
106.	2	1	2
107.	3	2	3
108.	1	1	1
109.	2	1	2
110.	3	2	3
111.	1	1	1
112.	2	1	2
113.	3	2	3
114.	1	1	2

图 13.22 对数据进行展示

在如图 13.22 所示的分析结果中可以看出，数据的总体质量还是可以的，没有极端异常值，

变量间的量纲差距也是可以接受的，可以进入下一步的分析。

图 13.23 是以 V1 为因变量，以 V2、V3 为自变量，进行最小二乘回归分析的结果。

```
. reg V1 V2 V3

      Source |       SS       df       MS              Number of obs =     114
-------------+------------------------------           F(  2,   111) =  112.42
       Model |  51.0694713     2  25.5347356           Prob > F      =  0.0000
    Residual |  25.2112305   111  .227128203           R-squared     =  0.6695
-------------+------------------------------           Adj R-squared =  0.6635
       Total |  76.2807018   113  .675050458           Root MSE      =  .47658

------------------------------------------------------------------------------
          V1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          V2 |   .7219745   .1069115     6.75   0.000     .5101221    .9338268
          V3 |   .5331441   .0665703     8.01   0.000     .4012307    .6650576
       _cons |  -.1616663    .144898    -1.12   0.267    -.4487914    .1254587
------------------------------------------------------------------------------
```

图 13.23　最小二乘回归分析

从上述分析结果中可以看出共有 114 个样本参与了分析，模型的 F 值(2, 111) =112.42，P 值（Prob > F） = 0.0000，说明模型整体上比较显著。模型的可决系数（R-squared）为 0.6695，模型修正的可决系数（Adj R-squared）为 0.6635，说明模型的解释能力差强人意。

变量 V2 的系数标准误是 0.1069115，t 值为 6.75，P 值为 0.000，系数显著性是非常不错的，95%的置信区间为[0.5101221, 0.9338268]。变量 V3 的系数标准误是 0.0665703，t 值为 8.01，P 值为 0.000，系数是非常显著的，95%的置信区间为[0.4012307, 0.6650576]。常数项的系数标准误是 0.144898，t 值为-1.12，P 值为 0.267，系数显著性是勉强过得去的，95%的置信区间为[-0.4487914, 0.1254587]。

从上述分析结果可以得到最小二乘模型的回归方程是：

V1=0.7219745*V2 +0.5331441*V3 -0.1616663

从上面的分析可以看出最小二乘线性模型的整体显著性、系数显著性以及模型的整体解释能力都是可以的。

图 13.24 是以 V1 为因变量，以 V2、V3 为自变量，进行有序 Logistic 回归分析的结果。

```
. ologit  V1 V2 V3

Iteration 0:   log likelihood = -123.89889
Iteration 1:   log likelihood = -70.339544
Iteration 2:   log likelihood = -67.506639
Iteration 3:   log likelihood = -67.476317
Iteration 4:   log likelihood = -67.476268
Iteration 5:   log likelihood = -67.476268

Ordered logistic regression                       Number of obs   =        114
                                                  LR chi2(2)      =     112.85
                                                  Prob > chi2     =     0.0000
Log likelihood = -67.476268                       Pseudo R2       =     0.4554

------------------------------------------------------------------------------
          V1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          V2 |   2.830534   .5714954     4.95   0.000     1.710424    3.950645
          V3 |   2.357495   .4023518     5.86   0.000       1.5689     3.14609
-------------+----------------------------------------------------------------
       /cut1 |    7.23271   1.018615                      5.236262    9.229158
       /cut2 |   10.36945   1.340869                      7.741391     12.9975
------------------------------------------------------------------------------
```

图 13.24　进行有序 Logistic 回归分析

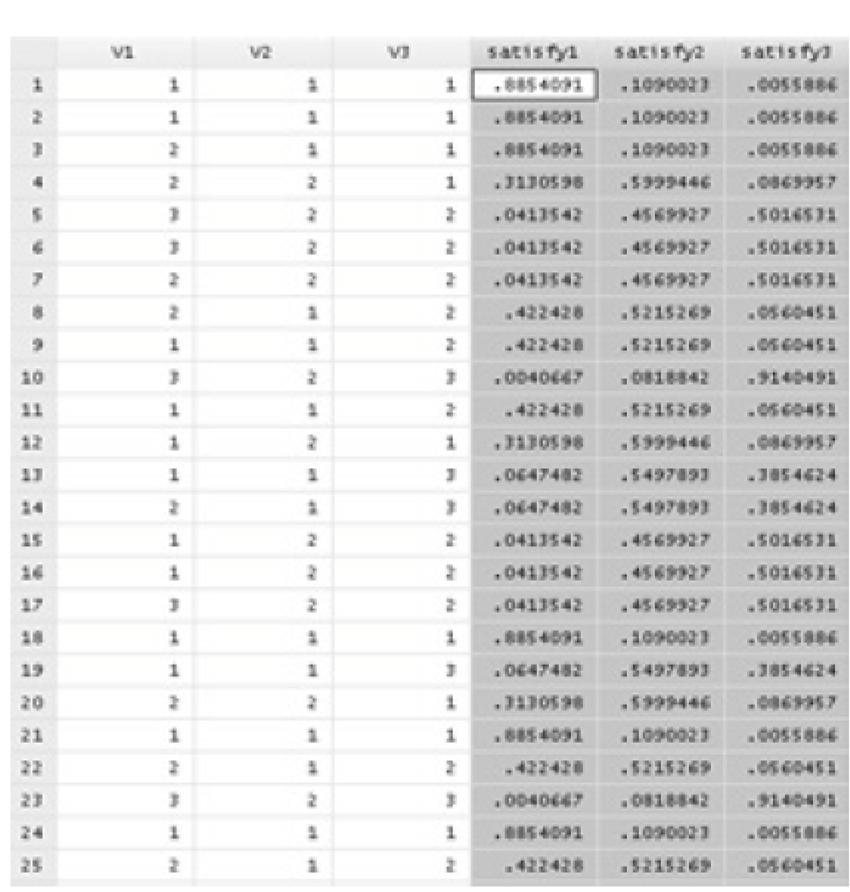
从图 13.24 可以看出有序 Logistic 模型与最小二乘回归估计效果相差不大。模型的整体显

著性 P 值远远低于 5%（Prob > chi2=0.0079）。伪 R 方达到 45.54% （Pseudo R2 = 0.4554）。

从图 13.24 中可以看到 V2 和 V3 系数在第 2 组和第 3 组都是大于 0 的，这意味着 V2 和 V3 两个变量的值越大越容易被分到后面的组，表示性别为女，学历越高，越容易被分到消费者满意程度较低的组。

“/cut1”和“/cut2”表示的含义是割点的估计值，两个割点把样本分成了 3 个区间，也就是消费者 3 个不同的满意程度。当样本的因变量拟合值在“/cut1”之下时，它被分到第 1 组，消费者满意程度为最高；当样本的因变量拟合值在“/cut1”之上且在“/cut2”之下时，它被分到第 2 组，消费者满意程度为中度；当样本的因变量拟合值在“/cut2”之上时，它被分到第 3 组，消费者满意程度为最低。

图 13.25 是根据模型预测每个观测样本消费者满意程度的可能性的结果。选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 13.25 所示的 satisfy1~satisfy3 数据。



	V1	V2	V3	satisfy1	satisfy2	satisfy3
1	1	1	1	.8854091	.1090023	.0055886
2	1	1	1	.8854091	.1090023	.0055886
3	2	1	1	.8854091	.1090023	.0055886
4	2	2	1	.3130598	.5999446	.0869957
5	3	2	2	.0413542	.4569927	.5016531
6	3	2	2	.0413542	.4569927	.5016531
7	2	2	2	.0413542	.4569927	.5016531
8	2	1	2	.422428	.5215269	.0560451
9	1	1	2	.422428	.5215269	.0560451
10	3	2	3	.0040667	.0818842	.9140491
11	1	1	2	.422428	.5215269	.0560451
12	1	2	1	.3130598	.5999446	.0869957
13	1	1	3	.0647482	.5497893	.3854624
14	2	1	3	.0647482	.5497893	.3854624
15	1	2	2	.0413542	.4569927	.5016531
16	1	2	2	.0413542	.4569927	.5016531
17	3	2	2	.0413542	.4569927	.5016531
18	1	1	1	.8854091	.1090023	.0055886
19	1	1	3	.0647482	.5497893	.3854624
20	2	2	1	.3130598	.5999446	.0869957
21	1	1	1	.8854091	.1090023	.0055886
22	2	1	2	.422428	.5215269	.0560451
23	3	2	3	.0040667	.0818842	.9140491
24	1	1	1	.8854091	.1090023	.0055886
25	2	1	2	.422428	.5215269	.0560451

图 13.25　根据模型预测消费者满意程度

如图 13.25 所示，第 1 个观测样本为男性，学历为大学专科及以下，他有 88%以上的概率进入第 1 组，即消费者满意程度为最高，有极小的甚至可以忽略不计的概率被分到第 3 组，即消费者满意程度为最低。其他的观测样本，读者可以按照类似的方法逐一进行分析，可以看出，我们的模型构建的不错，模型的预测能力也是比较优秀的。

13.3.5　案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

延伸：使用 Probit 模型对有序分类因变量进行拟合

以本节中介绍的实例进行说明，那么操作命令如下。

（1）oprobit V1 V2 V3

本命令的含义是以 V1 为因变量，以 V2、V3 为自变量，进行 Probit 回归分析，研究变量

之间的因果影响关系。

（2）predict satisfy1 satisfy2 satisfy3

本命令旨在估计因变量的拟合值。它创建一个命名为 yhat 的新变量，等于最近一次 Probit 模型基础上 y=1 的预测概率。

在命令窗口输入命令并按回车键进行确认，结果如图 13.26 和图 13.27 所示。

图 13.26 是以 V1 为因变量，以 V2、V3 为自变量，进行有序 Probit 回归分析的结果。

```
. oprobit  V1 V2 V3

Iteration 0:   log likelihood = -123.89889
Iteration 1:   log likelihood = -68.713098
Iteration 2:   log likelihood = -68.043379
Iteration 3:   log likelihood = -68.039793
Iteration 4:   log likelihood = -68.039793

Ordered probit regression                         Number of obs   =        114
                                                  LR chi2(2)      =     111.72
                                                  Prob > chi2     =     0.0000
Log likelihood = -68.039793                       Pseudo R2       =     0.4508

------------------------------------------------------------------------------
          V1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          V2 |   1.593445   .2938913     5.42   0.000     1.017428    2.169461
          V3 |   1.322247   .2041027     6.48   0.000     .9222136    1.722281
-------------+----------------------------------------------------------------
       /cut1 |    4.16763   .5212681                      3.145963    5.189297
       /cut2 |   5.874543   .6518366                      4.596967    7.152119
------------------------------------------------------------------------------
```

图 13.26　进行有序 Probit 回归分析

从上面的分析结果中可以看出，Probit 模型与 Logistic 模型所得结果相差不大，对本结果的详细解读限于篇幅，不再赘述。

图 13.27 是根据模型预测每个观测样本消费者满意程度的可能性的结果。选择“Data” | “Data Editor” | “Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 13.27 所示的 satisfy1~satisfy3 数据。

	V1	V2	V3	satisfy1	satisfy2	satisfy3
1	1	1	1	.8947037	.1037524	.0015439
2	1	1	1	.8947037	.1037524	.0015439
3	2	1	1	.8947037	.1037524	.0015439
4	2	2	1	.3663609	.5475764	.0860628
5	3	2	2	.0480807	.4691316	.4827877
6	3	2	2	.0480807	.4691316	.4827877
7	2	2	2	.0480807	.4691316	.4827877
8	2	1	2	.4719735	.4771698	.0508567
9	1	1	2	.4719735	.4771698	.0508567
10	3	2	3	.0014133	.0990196	.8995672
11	1	1	2	.4719735	.4771698	.0508567
12	1	2	1	.3663609	.5475764	.0860628
13	1	1	3	.0818768	.5414977	.3766255
14	2	1	3	.0818768	.5414977	.3766255
15	1	2	2	.0480807	.4691316	.4827877
16	1	2	2	.0480807	.4691316	.4827877
17	3	2	2	.0480807	.4691316	.4827877
18	1	1	1	.8947037	.1037524	.0015439
19	1	1	3	.0818768	.5414977	.3766255
20	2	2	1	.3663609	.5475764	.0860628
21	1	1	1	.8947037	.1037524	.0015439
22	2	1	2	.4719735	.4771698	.0508567
23	3	2	3	.0014133	.0990196	.8995672
24	1	1	1	.8947037	.1037524	.0015439
25	2	1	2	.4719735	.4771698	.0508567

图 13.27　根据模型预测消费者满意程度

如图 13.27 所示，第 1 个观测样本为男性，学历为大学专科及以下，他有 89%以上的概率进入第 1 组，即消费者满意程度为最高，有极小的甚至可以忽略不计的概率被分到第 3 组，即消费者满意程度为最低。其他的观测样本，读者可以按照类似的方法逐一进行分析，可以看出，

我们的模型构建的不错，模型的预测能力也是比较优秀的。

13.4 本章习题

（1）表 13.4 给出了 20 名前列腺癌患者的相关数据。试用二元 Logistic 回归方法分析患者前列腺细胞癌转移情况（有转移 y=1、无转移 y=0）与患者年龄、前列腺细胞癌血管内皮生长因子（由低到高共 3 个等级）、术前探针活检病理分级（从低到高共 4 级）、酸性磷酸酯酶、前列腺细胞癌分期（由低到高共 4 期 ）之间的关系。

表 13.4 20 名前列腺癌患者的相关数据

编号	前列腺细胞癌转移情况	年龄	前列腺细胞癌血管内皮生长因子	术前探针活检病理分级	酸性磷酸酯酶/个/μL	前列腺细胞癌分期
1	0	66	3	3	46	1
2	1	45	2	2	60	2
3	1	79	1	1	50	3
4	0	65	2	3	50	2
5	0	55	3	4	60	3
6	0	58	3	3	43	2
7	1	43	1	2	70	1
8	0	45	2	4	56	4
9	0	51	1	1	76	1
10	1	57	3	1	70	2
11	0	66	2	3	50	1
12	1	30	3	4	55	3
13	0	53	1	1	59	1
14	0	34	3	2	49	2
15	1	38	1	4	35	3
16	0	41	1	2	67	1
17	0	16	1	3	68	1
18	1	34	3	2	67	3
19	1	46	1	2	51	3
20	0	72	3	4	72	2

（2）表 13.5 给出了山东省某医院 20 名听力低下患者听力监测结果的数据。试用多元 Logistic 回归方法分析听力低下程度（由轻到重共 3 级）与年龄、性别（1 代表男性，2 代表女性）之间的关系。

表 13.5 山东省某医院 20 名听力低下患者听力监测结果的数据

编号	听力低下程度	性别	年龄
1	1	1	55
2	3	2	55
3	2	1	54
4	2	2	66
5	3	2	76
6	2	2	47

（续表）

编号	听力低下程度	性别	年龄
7	2	2	67
8	2	1	58
9	1	1	34
10	3	2	28
11	3	1	67
12	2	2	67
13	3	1	75
14	2	1	48
15	1	2	55
16	3	2	75
17	3	2	47
18	1	1	55
19	1	1	65
20	3	2	76

（3）某公司 114 位员工 2012 年的绩效考核情况的相关信息如表 13.6 所示。试用有序 Logistic 回归方法分析员工绩效考核情况（1 表示非常优秀，2 表示基本可以，3 表示不过关）与性别（1 代表男性，2 代表女性）、级别（1 表示高级员工，2 表示中级员工，3 表示初级员工）之间的关系。

表 13.6 某公司 114 位员工绩效考核情况数据

编号	绩效考核情况	性别	级别
1	3	2	3
2	1	1	2
3	1	2	1
4	1	1	3
5	2	1	3
6	1	2	2
…	…	…	…
109	2	1	2
110	2	1	3
111	1	2	2
112	2	2	2
113	2	1	2
114	1	1	2

第 14 章　Stata 因变量受限回归分析

前面我们讲述的回归分析方法都要求因变量或连续或离散。但是很多时候因变量观测样本数据会受到各种各样的限制，只能观测到满足一定条件的样本。例如，我们在统计某地区游客量时可能仅仅能够统计到知名景点，或者说游客人数大于某一特定值的景点游客量，又例如在统计工人的劳动时间时，失业工人的劳动时间一定只取 0，而不论失业的程度有多大有多深。根据因变量的受限特征，常用的因变量受限回归分析方法有两种，包括断尾回归分析和截取回归分析等。下面就以实例的方式一一介绍这两种方法在 Stata 中的应用。

14.1　实例一——断尾回归分析

14.1.1　断尾回归分析的功能与意义

断尾回归分析是针对因变量只有大于一定数值或者小于一定数值时才能被观测到的一种回归分析方法。或者说，因变量的取值范围是受到限制的，是不可能取到范围之外的数值的，通过一般的最小二乘回归分析得到的结论是不完美的。举例来说，如果研究某单位的薪酬情况，把年薪作为因变量，那么该因变量的取值范围就是大于 0 的，低于 0 是不可能的，是没有意义的。下面就介绍一下断尾回归分析在实例中的具体应用。

14.1.2　相关数据来源

	下载资源:\video\chap14\…
	下载资源:\sample\chap14\案例14.1.dta

【例 14.1】表 14.1 给出了某单位 88 名在岗职工的工龄、职称级别、月工作时间以及月工资收入情况。已知该单位的保底工资是 3000 元/月。试构建回归分析模型研究一下该单位职工的月工资收入受工龄、职称级别（1 表示初级职称，2 表示中级职称，3 表示高级职称）、月工作时间等变量的影响情况。

表 14.1　某单位 88 名在岗职工的工龄、职称级别、工作时间以及月工资情况数据

编号	月工资收入/元	月工作时间/小时	工龄/年	职称级别
1	6389	110	9	1
2	5327	108	8	1
3	4529	88	4	1

（续表）

编号	月工资收入/元	月工作时间/小时	工龄/年	职称级别
4	8723	135	10	2
5	10213	164	15	3
6	4596	86	6	1
…	…	…	…	…
83	8537	135	11	2
84	8123	120	10	2
85	7565	113	9	1
86	10330	165	16	3
87	7429	119	9	2
88	7625	123	9	2

14.1.3 Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有 4 个变量，分别是月工资收入、月工作时间、工龄以及职称级别。我们把月工资收入变量定义为 salary，把月工作时间变量定义为 hour，把工龄变量定义为 year，把职称级别变量定义为 grade。变量类型及长度采取系统默认方式，然后录入相关数据。相关操作我们在第 1 章中已有详细讲述。录入完成后数据如图 14.1 所示。

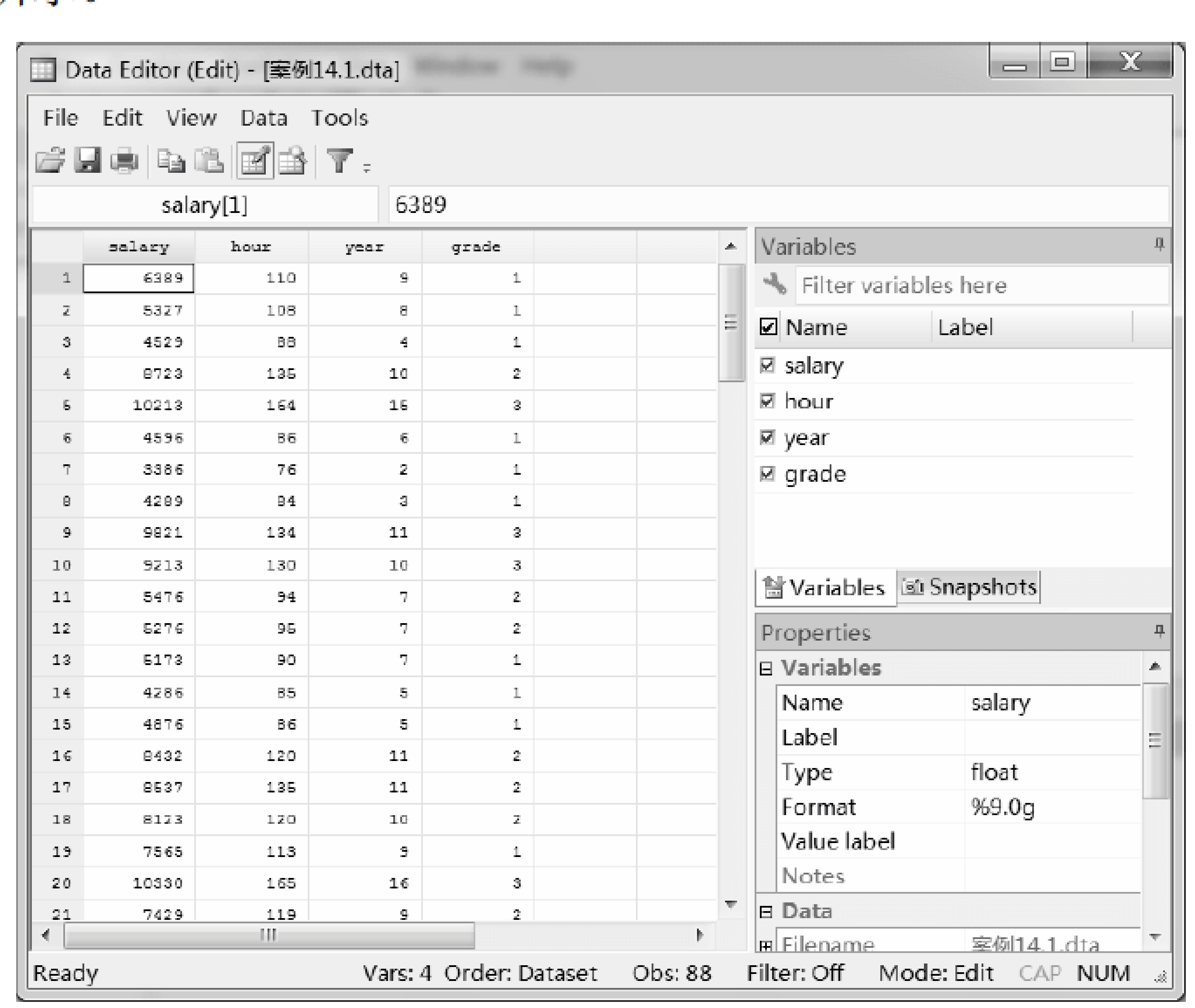


图 14.1　案例 14.1 数据

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入如下命令。

- list salary hour year grade：本命令的含义是对 4 个变量所包含的样本数据进行一一展示，以便简单直观地观测出数据的具体特征，为深入分析做好必要准备。

- reg salary hour year grade：本命令的含义是以 salary 为因变量，以 hour、year、grade 为自变量，进行最小二乘回归分析，研究变量之间的因果影响关系。
- truncreg salary hour year grade,ll(3000)：本命令的含义是以 salary 为因变量，以 hour、year、grade 为自变量，进行断尾回归分析，研究变量之间的因果影响关系。
- test hour year grade：本命令的含义是对断尾回归分析估计的各个自变量的系数进行假设检验，检验其显著程度。
- predict yhat：本命令的含义是估计因变量的拟合值。
- predict e,resid：本命令的含义是估计断尾回归分析的残差。

03 设置完毕后，按键盘上的回车键，等待输出结果。

14.1.4　结果分析

在 Stata 14.0 主界面的结果窗口可以看到如图 14.2~图 14.7 所示的分析结果。

图 14.2 是对数据进行展示的结果。它的目的是通过对变量所包含的样本数据进行一一展示，以便简单直观地观测出数据的具体特征，为深入分析做好必要准备。

```
. list   salary hour year grade
```

	salary	hour	year	grade
1.	6389	110	9	1
2.	5327	108	8	1
3.	4529	88	4	1
4.	8723	135	10	2
5.	10213	164	15	3
6.	4596	86	6	1
7.	3386	76	2	1
8.	4289	84	3	1
9.	9821	134	11	3
10.	9213	130	10	3
11.	5476	94	7	2
12.	5276	95	7	2
13.	5173	90	7	1
14.	4286	85	5	1
15.	4876	86	5	1
16.	8432	120	11	2
17.	8537	135	11	2
18.	8123	120	10	2
19.	7565	113	9	1
20.	10330	165	16	3
21.	7429	119	9	2
22.	7625	123	9	2
23.	6389	110	9	1
24.	5327	108	8	1
25.	4529	88	4	1
26.	8723	135	10	2
27.	10213	164	15	3
28.	4596	86	6	1
29.	3386	76	2	1
30.	4289	84	3	1
31.	9821	134	11	3
32.	9213	130	10	3
33.	5476	94	7	2
34.	5276	95	7	2
35.	5173	90	7	1
36.	4286	85	5	1
37.	4876	86	5	1
38.	8432	120	11	2
39.	8537	135	11	2
40.	8123	120	10	2
41.	7565	113	9	1
42.	10330	165	16	3
43.	7429	119	9	2
44.	7625	123	9	2
45.	6389	110	9	1
46.	5327	108	8	1
47.	4529	88	4	1
48.	8723	135	10	2
49.	10213	164	15	3
50.	4596	86	6	1
51.	3386	76	2	1
52.	4289	84	3	1
53.	9821	134	11	3
54.	9213	130	10	3
55.	5476	94	7	2
56.	5276	95	7	2
57.	5173	90	7	1
58.	4286	85	5	1
59.	4876	86	5	1
60.	8432	120	11	2
61.	8537	135	11	2
62.	8123	120	10	2
63.	7565	113	9	1
64.	10330	165	16	3
65.	7429	119	9	2
66.	7625	123	9	2
67.	6389	110	9	1
68.	5327	108	8	1
69.	4529	88	4	1
70.	8723	135	10	2
71.	10213	164	15	3
72.	4596	86	6	1
73.	3386	76	2	1
74.	4289	84	3	1
75.	9821	134	11	3
76.	9213	130	10	3
77.	5476	94	7	2
78.	5276	95	7	2
79.	5173	90	7	1
80.	4286	85	5	1
81.	4876	86	5	1
82.	8432	120	11	2
83.	8537	135	11	2
84.	8123	120	10	2
85.	7565	113	9	1
86.	10330	165	16	3
87.	7429	119	9	2
88.	7625	123	9	2

图 14.2　对数据进行展示

在如图 14.2 所示的分析结果中可以看出，数据的总体质量还是可以的，没有极端异常值，变量间的量纲差距也是可以接受的，可以进入下一步的分析。

图 14.3 是以 salary 为因变量，以 hour、year、grade 为自变量，进行最小二乘回归分析的结果。

```
. reg salary hour year grade

      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  3,    84) =  430.16
       Model |   371452125     3   123817375           Prob > F      =  0.0000
    Residual |  24178631.5    84  287840.851           R-squared     =  0.9389
-------------+------------------------------           Adj R-squared =  0.9367
       Total |   395630756    87  4547479.96           Root MSE      =  536.51

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        hour |   51.93677   9.024075     5.76   0.000      33.9914    69.88213
        year |   120.8774   59.99078     2.01   0.047      1.57913    240.1756
       grade |   572.1885   135.5076     4.22   0.000     302.7168    841.6602
       _cons |  -1006.138     491.17    -2.05   0.044    -1982.884     -29.393
------------------------------------------------------------------------------
```

图 14.3　最小二乘回归分析

从上述分析结果中可以看出共有 88 个样本参与了分析，模型的 F 值(3, 84) = 430.16，P 值（Prob > F） = 0.0000，说明模型整体上是很显著的。模型的可决系数（R-squared）为 0.9389，模型修正的可决系数（Adj R-squared）为 0.9367，说明模型的解释能力也是非常好的。

变量 hour 的系数标准误是 9.024075，t 值为 5.76，P 值为 0.000，系数是非常显著的，95%的置信区间为[33.9914, 69.88213]。变量 year 的系数标准误是 59.99078，t 值为 2.01，P 值为 0.047，系数是比较显著的，95%的置信区间为[1.57913, 240.1756]。变量 grade 的系数标准误是 135.5076，t 值为 4.22，P 值为 0.000，系数是非常显著的，95%的置信区间为[302.7168,841.6602]。常数项的系数标准误是 491.17，t 值为-2.05，P 值为 0.044，系数也是比较显著的，95%的置信区间为[-1982.884, -29.393]。

从上述分析结果可以得到最小二乘模型的回归方程：

salary=51.93677*hour+120.8774*year+572.1885*grade-1006.138

从上面的分析可以看出最小二乘线性模型的整体显著性、系数显著性以及模型的整体解释能力都很不错。结论是该单位工人的月工资都是与月工作时间、工龄、职称级别等呈显著正向变化的。

图 14.4 是以 salary 为因变量，以 hour、year、grade 为自变量，进行断尾回归分析的结果。其中断尾点设置的是 3000。

从图 14.4 可以看出断尾回归分析模型相对于最小二乘回归模型得到了很大程度的改进。模型中各个变量系数的显著程度也有不同程度的提高，限于篇幅不再赘述。

图 14.5 是对断尾回归分析估计的各个自变量的系数进行假设检验的结果。

```
. truncreg salary hour year grade,ll(3000)
(note: 0 obs. truncated)

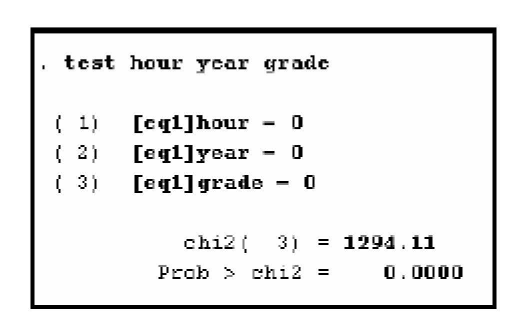
Fitting full model:

Iteration 0:   log likelihood = -675.57114
Iteration 1:   log likelihood = -675.52962
Iteration 2:   log likelihood = -675.52953
Iteration 3:   log likelihood = -675.52953

Truncated regression
Limit:   lower =       3000                      Number of obs =     88
         upper =       +inf                      Wald chi2(3)  =1294.11
Log likelihood = -675.52953                      Prob > chi2   = 0.0000

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        hour |   51.17762   8.973677     5.70   0.000     33.58954     68.7657
        year |   129.2596     60.534     2.14   0.033     10.61518    247.9041
       grade |   569.4026   133.5357     4.26   0.000     307.6774    831.1278
       _cons |  -991.5073   484.7357    -2.05   0.041    -1941.572   -41.44282
-------------+----------------------------------------------------------------
      /sigma |    528.295   40.62363    13.00   0.000     448.6741    607.9158
------------------------------------------------------------------------------
```

图 14.4　断尾回归分析



```
. test hour year grade

 ( 1)  [eq1]hour = 0
 ( 2)  [eq1]year = 0
 ( 3)  [eq1]grade = 0

           chi2(  3) = 1294.11
         Prob > chi2 =    0.0000
```

图 14.5　进行假设检验

从图 14.5 可以看出该模型非常显著，拟合很好。

图 14.6 是对因变量的拟合值的预测。

	salary	hour	year	grade	yhat
1	6389	110	9	1	6370.77
2	5327	108	8	1	6139.155
3	4529	88	4	1	4598.564
4	8723	135	10	2	8348.873
5	10213	164	15	3	11048.72
6	4596	86	6	1	4754.729
7	3386	76	2	1	3725.914
8	4289	84	3	1	4264.594
9	9821	134	11	3	8996.357
10	9213	130	10	3	8662.388
11	5476	94	7	2	5862.812
12	5276	95	7	2	5913.989
13	5173	90	7	1	5088.699
14	4286	85	5	1	4574.291
15	4876	86	5	1	4625.469
16	8432	120	11	2	7710.468
17	8537	135	11	2	8478.133
18	8123	120	10	2	7581.208
19	7565	113	9	1	6524.303
20	10330	165	16	3	11229.16
21	7429	119	9	2	7400.771
22	7625	123	9	2	7605.482
23	6389	110	9	1	6370.77
24	5327	108	8	1	6139.155
25	4529	88	4	1	4598.564
26	8723	135	10	2	8348.873
27	10213	164	15	3	11048.72
28	4596	86	6	1	4754.729

图 14.6　因变量的拟合值预测

关于因变量预测拟合值的意义我们在前面章节中已经论述过，此处旨在说明断尾回归也是可以进行预测拟合因变量值的，细节之处限于篇幅不再重复讲解。

图 14.7 是断尾回归分析得到的残差序列。

	salary	hour	year	grade	yhat	e
1	6389	110	9	1	6370.77	18.22976
2	5327	108	8	1	6139.155	-812.1553
3	4529	88	4	1	4598.564	-69.56438
4	8723	135	10	2	8348.873	374.127
5	10213	164	15	3	11048.72	-835.7247
6	4596	86	6	1	4754.729	-158.7284
7	3386	76	2	1	3725.914	-339.9137
8	4289	84	3	1	4264.594	24.40574
9	9821	134	11	3	8996.357	824.6425
10	9213	130	10	3	8662.388	550.6125
11	5476	94	7	2	5862.812	-386.8116
12	5276	95	7	2	5913.989	-637.9893
13	5173	90	7	1	5088.699	84.30145
14	4286	85	5	1	4574.291	-288.2912
15	4876	86	5	1	4625.469	250.5312
16	8432	120	11	2	7710.468	721.5317
17	8537	135	11	2	8478.133	58.86741
18	8123	120	10	2	7581.208	541.7913
19	7565	113	9	1	6524.303	1040.697
20	10330	165	16	3	11229.16	-899.162
21	7429	119	9	2	7400.771	28.22861
22	7625	123	9	2	7605.482	19.51814
23	6389	110	9	1	6370.77	18.22976
24	5327	108	8	1	6139.155	-812.1553
25	4529	88	4	1	4598.564	-69.56438
26	8723	135	10	2	8348.873	374.127
27	10213	164	15	3	11048.72	-835.7247
28	4596	86	6	1	4754.729	-158.7284

图 14.7　残差序列

14.1.5　案例延伸

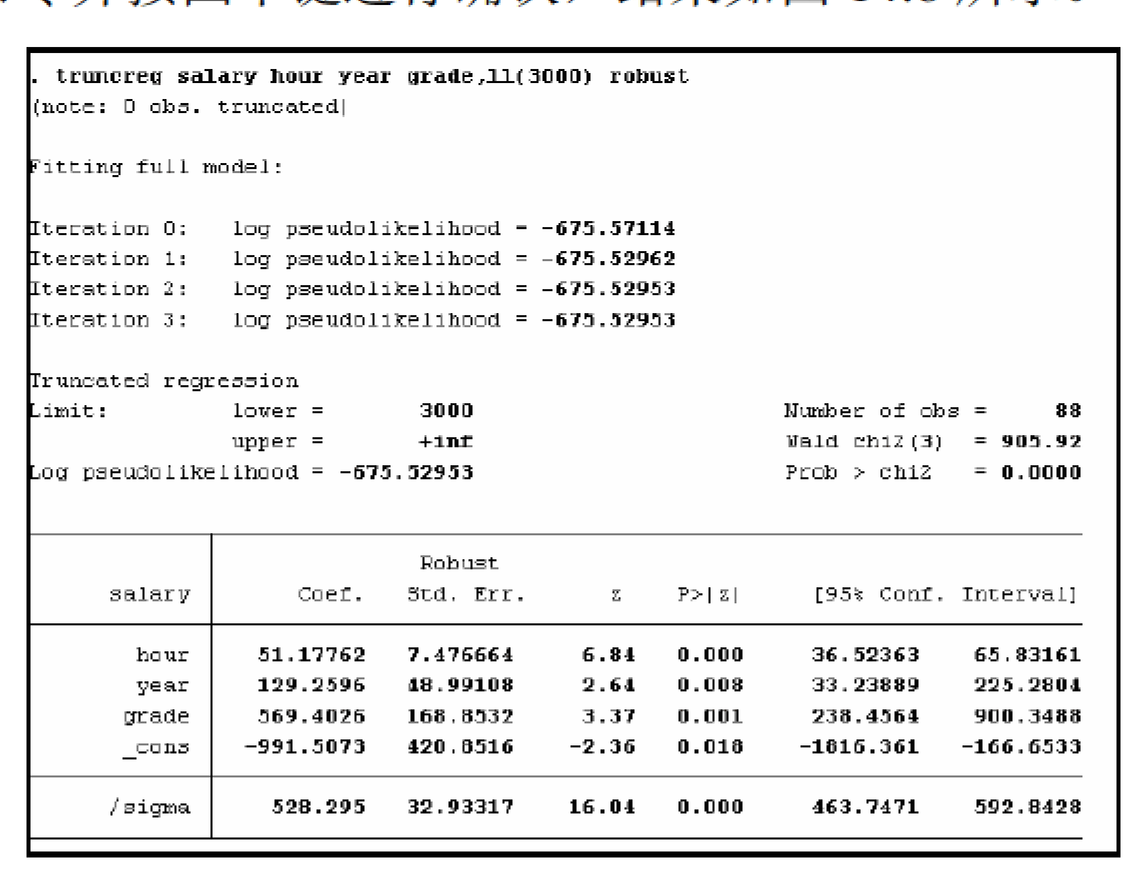
上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

延伸：使用稳健标准差进行断尾回归分析

与前面章节讲述的最小二乘回归分析类似，我们在断尾回归分析中也可以使用稳健的标准差，以克服可能会有的异方差的存在对模型的整体有效性带来的不利影响。以本节中提到的案例为例，操作命令就是：

```
truncreg salary hour year grade,ll(3000) robust
```

在命令窗口输入命令并按回车键进行确认，结果如图 14.8 所示。



```
. truncreg salary hour year grade,ll(3000) robust
(note: 0 obs. truncated)

Fitting full model:

Iteration 0:   log pseudolikelihood = -675.57114
Iteration 1:   log pseudolikelihood = -675.52962
Iteration 2:   log pseudolikelihood = -675.52953
Iteration 3:   log pseudolikelihood = -675.52953

Truncated regression
Limit:         lower =        3000                 Number of obs =      88
               upper =        +inf                 Wald chi2(3)  =  905.92
Log pseudolikelihood = -675.52953                  Prob > chi2   =  0.0000

                            Robust
      salary       Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]

        hour    51.17762    7.476664    6.84   0.000     36.52363    65.83161
        year    129.2596    48.99108    2.64   0.008     33.23889    225.2804
       grade    569.4026    168.8532    3.37   0.001     238.4564    900.3488
       _cons   -991.5073    420.8516   -2.36   0.018    -1816.361   -166.6533

      /sigma     528.295    32.93317   16.04   0.000     463.7471    592.8428
```

图 14.8　分析结果图

从上面的分析结果中可以看出模型中各变量的系数显著性较没有使用稳健标准差进行断尾回归分析时有了进一步的提高，模型更加完美。

14.2　实例二——截取回归分析

14.2.1　截取回归分析的功能与意义

截取回归分析是针对当因变量大于一定数值或者小于一定数值时仅能有一种取值时的回归分析方法。或者说，因变量的取值范围是受到限制的，当因变量大于一定值时，以后不管程度如何，统统被记录为某一特定值。在这种情况下，通过一般的最小二乘回归分析得到的结论是不完美的。举例来说，如果研究某单位的薪酬情况，该单位采取封顶薪酬方式，把年薪作为因变量，那么该因变量的取值范围就低于一定值。下面就介绍一下截取回归分析在实例中的具体应用。

14.2.2　相关数据来源

	下载资源:\video\chap14\…
	下载资源:\sample\chap14\案例14.2.dta

【例 14.2】表 14.2 给出了某单位 78 名在岗职工的工龄、职称级别、月工作时间以及月工资情况。已知该单位的封顶工资是 11000 元/月。试构建回归分析模型研究一下该单位职工的月工资受工龄、职称级别（1 表示初级职称，2 表示中级职称，3 表示高级职称）、月工作时间等变量的影响情况。

表 14.2　某单位 78 名在岗职工的工龄、职称级别、月工作时间以及月工资情况数据

编号	月工资收入/元	月工作时间/小时	工龄/年	职称级别
1	4596	86	6	1
2	3386	76	2	1
3	4289	84	3	1
4	9821	134	11	3
5	9213	130	10	3
6	5476	94	7	2
…	…	…	…	…
73	5276	95	7	2
74	5173	90	7	1
75	4286	85	5	1
76	4876	86	5	1
77	8432	120	11	2
78	8537	135	11	2

14.2.3　Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有 4 个变量，分别是月

工资收入、月工作时间、工龄以及职称级别。我们把月工资收入变量定义为 salary，把月工作时间变量定义为 hour，把工龄变量定义为 year，把职称级别变量定义为 grade。变量类型及长度采取系统默认方式，然后录入相关数据。相关操作在第 1 章中已有详细讲述。录入完成后数据如图 14.9 所示。

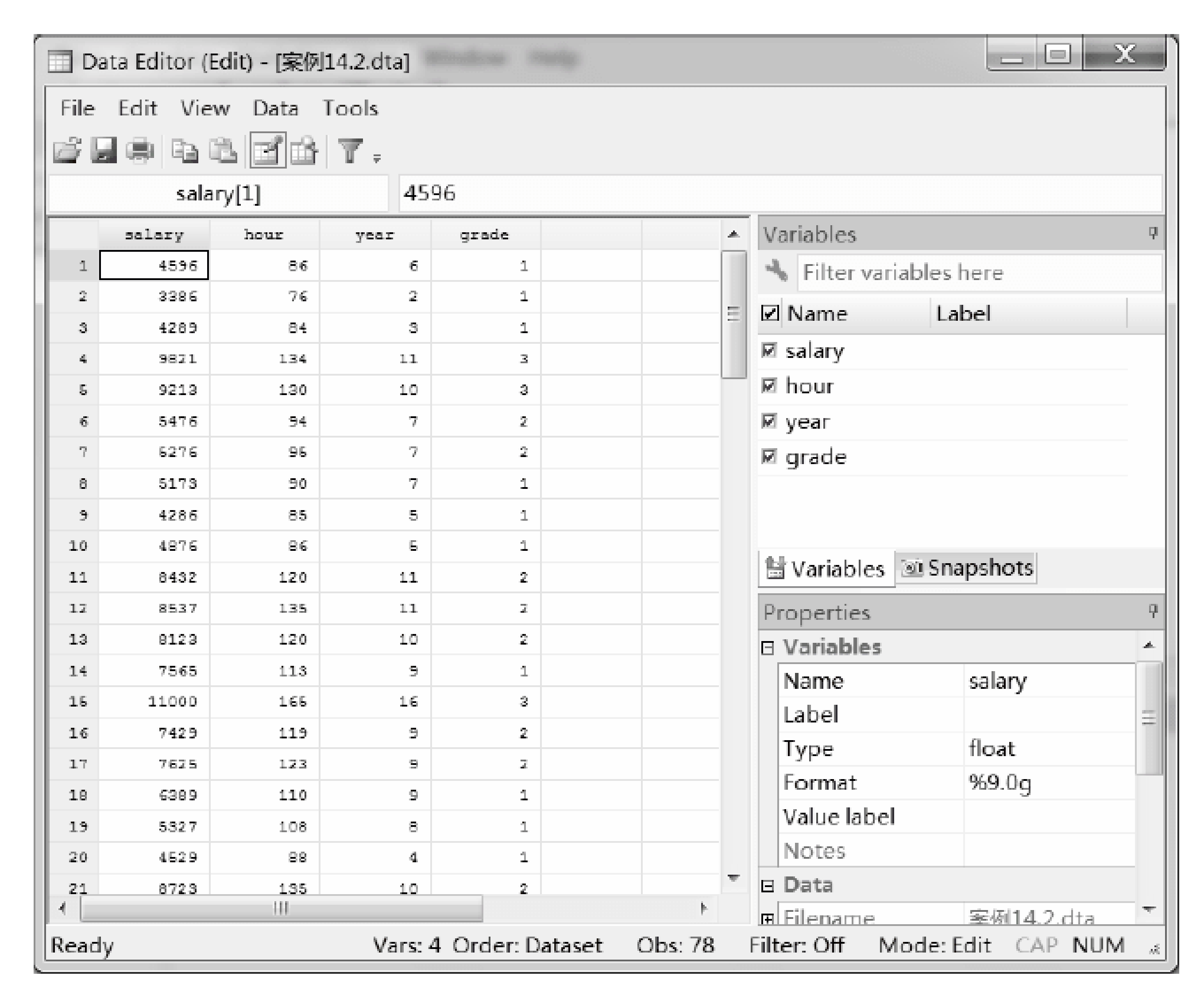


图 14.9　案例 14.2 数据

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的"Command"文本框中输入如下命令。

- list salary hour year grade：本命令的含义是对 4 个变量所包含的样本数据进行一一展示，以便简单直观地观测出数据的具体特征，为深入分析做好必要准备。
- reg salary hour year grade：本命令的含义是以 salary 为因变量，以 hour、year、grade 为自变量，进行最小二乘回归分析，研究变量之间的因果影响关系。
- tobit salary hour year grade,ul(11000)：本命令的含义是以 salary 为因变量，以 hour、year、grade 为自变量，进行断尾回归分析，研究变量之间的因果影响关系。
- test hour year grade：本命令的含义是对断尾回归分析估计的各个自变量的系数进行假设检验，检验其显著程度。
- predict yhat：本命令的含义是估计因变量的拟合值。

03 设置完毕后，按键盘上的回车键，等待输出结果。

14.2.4　结果分析

在 Stata 14.0 主界面的结果窗口可以看到如图 14.10~图 14.14 所示的分析结果。

图 14.10 是对数据进行展示的结果。它的目的是通过对变量所包含的样本数据进行一一展

示，以便简单直观地观测出数据的具体特征，为深入分析做好必要准备。

. list salary hour year grade

	salary	hour	year	grade
1.	4596	86	6	1
2.	3386	76	2	1
3.	4289	84	3	1
4.	9821	134	11	3
5.	9213	130	10	3
6.	5476	94	7	2
7.	5276	95	7	2
8.	5173	90	7	1
9.	4286	85	5	1
10.	4876	86	5	1
11.	8432	120	11	2
12.	8537	135	11	2
13.	8123	120	10	2
14.	7565	113	9	1
15.	11000	165	16	3
16.	7429	119	9	2
17.	7625	123	9	2
18.	6389	110	9	1
19.	5327	108	8	1
20.	4529	88	4	1
21.	8723	135	10	2
22.	11000	164	15	3
23.	4596	86	6	1
24.	3000	76	2	1
25.	3000	84	3	1
26.	9821	134	11	3
27.	9213	130	10	3
28.	5476	94	7	2
29.	5276	95	7	2
30.	5173	90	7	1
31.	4286	85	5	1
32.	4876	86	5	1
33.	8432	120	11	2
34.	8537	135	11	2
35.	8123	120	10	2
36.	7565	113	9	1
37.	11000	165	16	3
38.	7429	119	9	2
39.	7625	123	9	2
40.	6389	110	9	1
41.	5327	108	8	1
42.	3000	88	4	1
43.	8723	135	10	2
44.	11000	164	15	3
45.	4596	86	6	1
46.	3000	76	2	1
47.	4289	84	3	1
48.	9821	134	11	3
49.	9213	130	10	3
50.	5476	94	7	2
51.	5276	95	7	2
52.	5173	90	7	1
53.	4286	85	5	1
54.	4876	86	5	1
55.	8432	120	11	2
56.	8537	135	11	2
57.	8123	120	10	2
58.	7565	113	9	1
59.	11000	165	16	3
60.	7429	119	9	2
61.	7625	123	9	2
62.	6389	110	9	1
63.	5327	108	8	1
64.	4529	88	4	1
65.	8723	135	10	2
66.	11000	164	15	3
67.	4596	86	6	1
68.	3386	76	2	1
69.	4289	84	3	1
70.	11000	159	11	3
71.	9213	130	10	3
72.	5476	94	7	2
73.	5276	95	7	2
74.	5173	90	7	1
75.	4286	85	5	1
76.	4876	86	5	1
77.	8432	120	11	2
78.	8537	135	11	2

图 14.10　对数据进行展示

从图 14.10 所示的分析结果中可以看出，数据的总体质量还是可以的，没有极端异常值，变量间的量纲差距也是可以接受的，可以进入下一步的分析。

图 14.11 是以 salary 为因变量，以 hour、year、grade 为自变量，进行最小二乘回归分析的结果。

. reg salary hour year grade

Source	SS	df	MS		
				Number of obs =	78
				F(3, 74) =	575.78
Model	404115911	3	134705304	Prob > F =	0.0000
Residual	17312650.2	74	233954.732	R-squared =	0.9589
				Adj R-squared =	0.9573
Total	421428561	77	5473098.19	Root MSE =	483.69

salary	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hour	53.02997	7.845277	6.76	0.000	37.39791	68.66203
year	182.4601	52.15133	3.50	0.001	78.54635	286.3739
grade	554.3572	131.2952	4.22	0.000	292.7458	815.9686
_cons	-1582.902	424.996	-3.72	0.000	-2429.725	-736.0785

图 14.11　最小二乘回归分析

从上述分析结果中可以看出共有 78 个样本参与了分析，模型的 F 值(3, 74) = 575.78，P 值（Prob > F）= 0.0000，说明模型整体上是很显著的。模型的可决系数（R-squared）为 0.9589，模型修正的可决系数（Adj R-squared）为 0.9573，说明模型的解释能力也是非常好的。

变量 hour 的系数标准误是 7.845277，t 值为 6.76，P 值为 0.000，系数是非常显著的，95%的置信区间为[37.39791, 68.66203]。变量 year 的系数标准误是 52.15133，t 值为 3.50，P 值为

0.001，系数是非常显著的，95%的置信区间为[78.54635, 286.3739]。变量 grade 的系数标准误是 131.2952，t 值为 4.22，P 值为 0.000，系数是非常显著的，95%的置信区间为[292.7458, 815.9686]。常数项的系数标准误是 424.996，t 值为-3.72，P 值为 0.000，系数也是比较显著的，95%的置信区间为[-2429.725, -736.0785]。

从上述分析结果可以得到最小二乘模型的回归方程是：

salary= 53.02997 *hour+ 182.4601 *year+ 554.3572 *grade -1582.902

从上面的分析可以看出最小二乘线性模型的整体显著性、系数显著性以及模型的整体解释能力都很不错。我们得到的结论是该单位工人的月工资是与其月工作时间、工龄、职称级别等呈显著正向变化的。

图 14.12 是以 salary 为因变量，以 hour、year、grade 为自变量，进行截取回归分析的结果。其中，截取上限设置的是 11000。

```
. tobit  salary hour year grade,ul(11000)

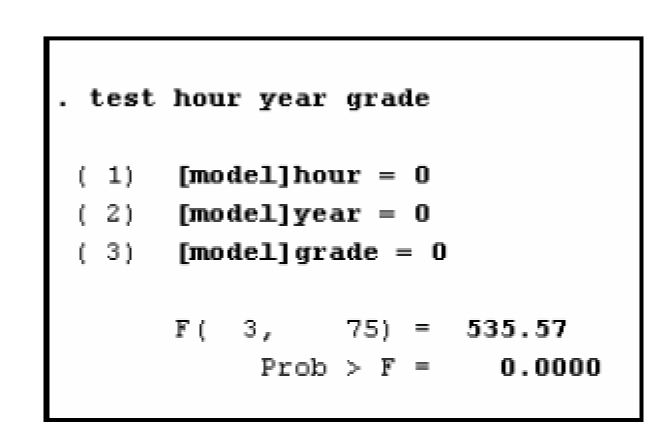
Tobit regression                                  Number of obs   =         78
                                                  LR chi2(3)      =     269.28
                                                  Prob > chi2     =     0.0000
Log likelihood = -531.46024                       Pseudo R2       =     0.2021

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        hour |   58.72234   7.167127     8.19   0.000     44.44469    72.99999
        year |   207.5801   47.64429     4.36   0.000     112.6678    302.4924
       grade |   525.3432   115.7347     4.54   0.000     294.7878    755.8987
       _cons |  -2272.016   404.3246    -5.62   0.000    -3077.472    -1466.56
-------------+----------------------------------------------------------------
      /sigma |   425.8502   35.61834                      354.8948    496.8056
------------------------------------------------------------------------------
  Obs. summary:          0  left-censored observations
                        71     uncensored observations
                         7 right-censored observations at salary>=11000
```

图 14.12　截取回归分析结果图

从图 14.12 可以看出截取回归分析模型相对于最小二乘回归模型得到了很大程度的改进。模型中各个变量系数的显著程度也有不同程度的提高，限于篇幅不再赘述。

图 14.13 是对截取回归分析估计的各个自变量的系数进行假设检验的结果。



```
. test hour year grade

 ( 1)  [model]hour = 0
 ( 2)  [model]year = 0
 ( 3)  [model]grade = 0

       F(  3,    75) =  535.57
            Prob > F =    0.0000
```

图 14.13　进行假设检验

从图 14.13 可以看出该模型非常显著，拟合很好。

图 14.14 是对因变量的拟合值的预测。

	salary	hour	year	grade	yhat
1	4596	86	6	1	4548.929
2	3386	76	2	1	3131.385
3	4289	84	3	1	3808.744
4	9821	134	11	3	9456.188
5	9213	130	10	3	9013.719
6	5476	94	7	2	5751.631
7	5276	95	7	2	5810.354
8	5173	90	7	1	4991.398
9	4286	85	5	1	4282.626
10	4876	86	5	1	4341.349
11	8432	120	11	2	8108.732
12	8537	135	11	2	8989.567
13	8123	120	10	2	7901.152
14	7565	113	9	1	6757.172
15	11000	165	16	3	12314.48
16	7429	119	9	2	7634.85
17	7625	123	9	2	7869.739
18	6389	110	9	1	6581.005
19	5327	108	8	1	6255.98
20	4529	88	4	1	4251.213
21	8723	135	10	2	8781.987
22	11000	164	15	3	12048.18
23	4596	86	6	1	4548.929
24	3000	76	2	1	3131.385
25	3000	84	3	1	3808.744
26	9821	134	11	3	9456.188
27	9213	130	10	3	9013.719
28	5476	94	7	2	5751.631

图 14.14　查看数据

关于因变量预测拟合值的意义在前面章节已经论述了，此处旨在说明截取回归也是可以进行预测拟合因变量值的，细节之处限于篇幅不再重复讲解。

14.2.5　案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

1．延伸 1：使用稳健标准差进行截取回归分析

与前面章节讲述的最小二乘回归分析类似，在截取回归分析中也可以使用稳健的标准差，以克服可能会有的异方差的存在对模型的整体有效性带来的不利影响。以本节中提到的案例为例，操作命令就是：

```
tobit  salary hour year grade,ul(11000) robust
```

在命令窗口输入命令并按回车键进行确认，结果如图 14.15 所示。

```
. tobit  salary hour year grade,ul(11000) robust

Tobit regression                                Number of obs   =         78
                                                F(   3,      75) =     770.49
                                                Prob > F        =     0.0000
Log pseudolikelihood = -531.46024               Pseudo R2       =     0.2021

------------------------------------------------------------------------------
             |               Robust
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        hour |   58.72234   6.686075     8.78   0.000     45.40299    72.04168
        year |   207.5801   45.05987     4.61   0.000     117.8162     297.344
       grade |   525.3432   139.3285     3.77   0.000     247.7866    802.8998
       _cons |  -2272.016   331.9053    -6.84   0.000    -2933.365   -1610.667
-------------+----------------------------------------------------------------
      /sigma |   425.8502   37.64919                      350.8492    500.8513
------------------------------------------------------------------------------
  Obs. summary:          0  left-censored observations
                        71     uncensored observations
                         7 right-censored observations at salary>=11000
```

图 14.15　使用稳健标准差进行截取回归分析

从上面的分析结果中可以看出模型中各变量的系数显著性较没有使用稳健标准差进行截取回归分析时有了进一步的提高，模型更加完美。

2．延伸 2：设置下限进行截取回归分析

与设置上限类似，也可以设置截取回归的下限进行分析。以本节中提到的案例为例，如果设置保底工资为 3000，而不设置封顶工资，那么操作命令就是：

```
tobit  salary hour year grade,ll(3000)
```

在命令窗口输入命令并按回车键进行确认，结果如图 14.16 所示。

```
. tobit  salary hour year grade,ll(3000)

Tobit regression                                  Number of obs   =         78
                                                  LR chi2(3)      =     236.73
                                                  Prob > chi2     =     0.0000
Log likelihood = -568.55468                       Pseudo R2       =     0.1723

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        hour |   51.33354   8.021806     6.40   0.000     35.35329     67.3138
        year |   200.7987   53.76625     3.73   0.000     93.69078    307.9065
       grade |   552.1327    133.658     4.13   0.000     285.8723    818.3932
       _cons |  -1553.493   432.7089    -3.59   0.001    -2415.493   -691.4923
-------------+----------------------------------------------------------------
      /sigma |   492.2026   41.01325                         410.5    573.9051
------------------------------------------------------------------------------
  Obs. summary:          4  left-censored observations at salary<=3000
                        74     uncensored observations
                         0 right-censored observations
```

图 14.16　设置下限进行截取回归分析

模型结果的解读方式与前面所述类似，此处限于篇幅不再赘述。

3．延伸 3：同时设置上限和下限进行截取回归分析

以本节中提到的案例为例，如果设置保底工资为 3000，同时设置封顶工资为 11000，那么操作命令就是：

```
tobit  salary hour year grade,ll(3000) ul(11000)
```

在命令窗口输入命令并按回车键进行确认，结果如图 14.17 所示。

```
. tobit  salary hour year grade,ll(3000) ul(11000)

Tobit regression                                  Number of obs   =         78
                                                  LR chi2(3)      =     256.61
                                                  Prob > chi2     =     0.0000
Log likelihood = -508.94234                       Pseudo R2       =     0.2013

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        hour |   57.14519   7.517156     7.60   0.000     42.17025    72.12013
        year |   228.6658   50.69766     4.51   0.000     127.6709    329.6607
       grade |   520.8632   121.0532     4.30   0.000     279.7128    762.0137
       _cons |  -2270.666   422.2223    -5.38   0.000    -3111.776   -1429.556
-------------+----------------------------------------------------------------
      /sigma |   445.1417   38.79441                      367.8593    522.4241
------------------------------------------------------------------------------
  Obs. summary:          4  left-censored observations at salary<=3000
                        67     uncensored observations
                         7 right-censored observations at salary>=11000
```

图 14.17　同时设置上限和下限进行截取回归分析

模型结果的解读方式与前面所述类似，此处限于篇幅不再赘述。

14.3 本章习题

（1）表 14.3 给出了某医院 70 名在岗医生的从业年限、职称、诊疗人数以及满意度得分情况。已知所有医生的保底得分是 30 分。试构建回归分析模型研究一下该单位医生的满意度得分受从业年限、职称级别（1 表示初级职称，2 表示中级职称，3 表示高级职称）、诊疗人数等变量的影响情况。

表 14.3 某单位 70 名在岗医生的从业年限、职称级别、诊疗人数以及满意度得分情况数据

编号	满意度得分	诊疗人数	从业年限	职称级别
1	54.76	94	7	2
2	52.76	95	7	2
3	51.73	90	7	1
4	42.86	85	5	1
5	48.76	86	5	1
6	84.32	120	11	2
…	…	…	…	…
65	98.21	134	11	3
66	92.13	130	10	3
67	54.76	94	7	2
68	52.76	95	7	2
69	51.73	90	7	1
70	42.86	85	5	1

（2）表 14.4 给出了某地区 60 个旅游景点的游客量、投资金额、建成年限以及国家评级情况。已知该地区各景点的封顶接待量是 11000 人/次。试构建回归分析模型研究一下该地区 60 个旅游景点的游客量受投资金额、建成年限以及国家评级情况（1 表示 AA 级，2 表示 AAA 级，3 表示 AAAA 级）等变量的影响情况。

表 14.4 某地区 60 个旅游景点的游客量、投资金额、建成年限以及国家评级情况数据

编号	游客量/人/次	投资金额/万元	建成年限/年	国家评级情况
1	5276	95	7	2
2	5173	90	7	1
3	4286	85	5	1
4	4876	86	5	1
5	8432	120	11	2
6	8537	135	11	2
…	…	…	…	…
55	7625	123	9	2
56	6389	110	9	1
57	5327	108	8	1
58	4529	88	4	1
59	8723	135	10	2
60	11000	164	15	3

第 15 章　Stata 时间序列分析

时间序列分析是一种动态数据处理的统计方法。该方法基于随机过程理论和数理统计学方法，研究随机数据序列所遵从的统计规律，以此来解决实际问题。时间序列是随时间而变化、具有动态性和随机性的数字序列。在现实生活中，许多统计资料都是按照时间进行观测记录的，因此时间序列分析在实际分析中具有广泛的应用。

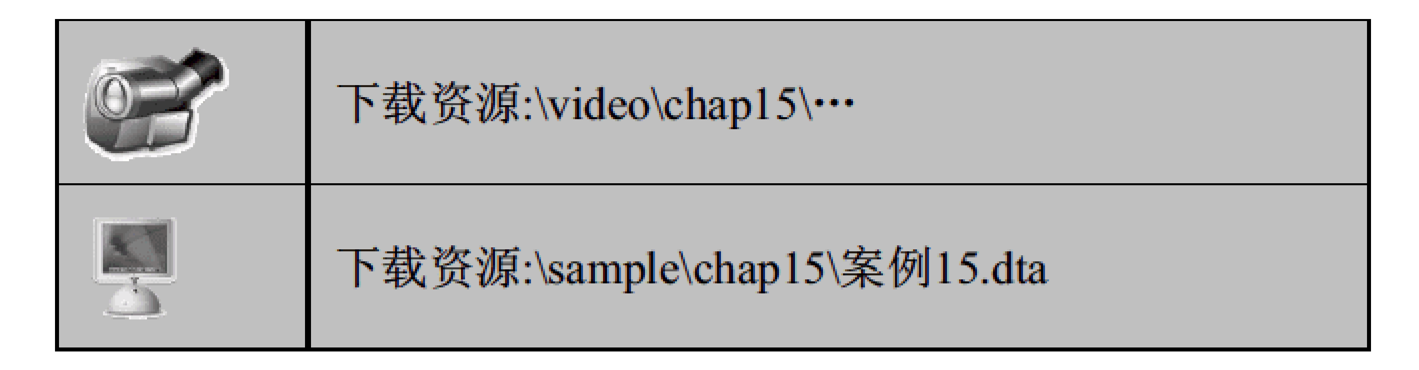
时间序列模型不同于一般的经济计量模型，其不以经济理论为依据，而是依据变量自身的变化规律，利用外推机制描述时间序列的变化。时间序列模型在处理的过程中必须明确考虑时间序列的非平稳性。本章我们就来对 Stata 中提供的时间序列分析功能进行一系列的实例分析。

15.1　时间序列分析的基本操作

15.1.1　时间序列分析的基本操作概述

在进行时间序列分析前，我们往往需要对数据进行预处理。首先要分析的是该数据是否适合用时间序列分析，这往往需要我们提前对数据进行简单回归，然后再进行时间序列分析的基本操作，包括定义时间序列、绘制时间序列趋势图等。对于一个带有日期变量的数据文件，Stata 14.0 并不会自动识别并判定出该数据是否是时间序列数据，尤其是数据含有多个日期变量的情形，所以要选取出恰当的日期变量，然后定义时间序列。而绘制时间序列趋势图的意义是不言而喻的，通过该步操作我们可以迅速看出数据的变化特征，为后续更加精确地判断或者选择合适的模型做好必要准备。

15.1.2　相关数据来源



	下载资源:\video\chap15\…
	下载资源:\sample\chap15\案例15.dta

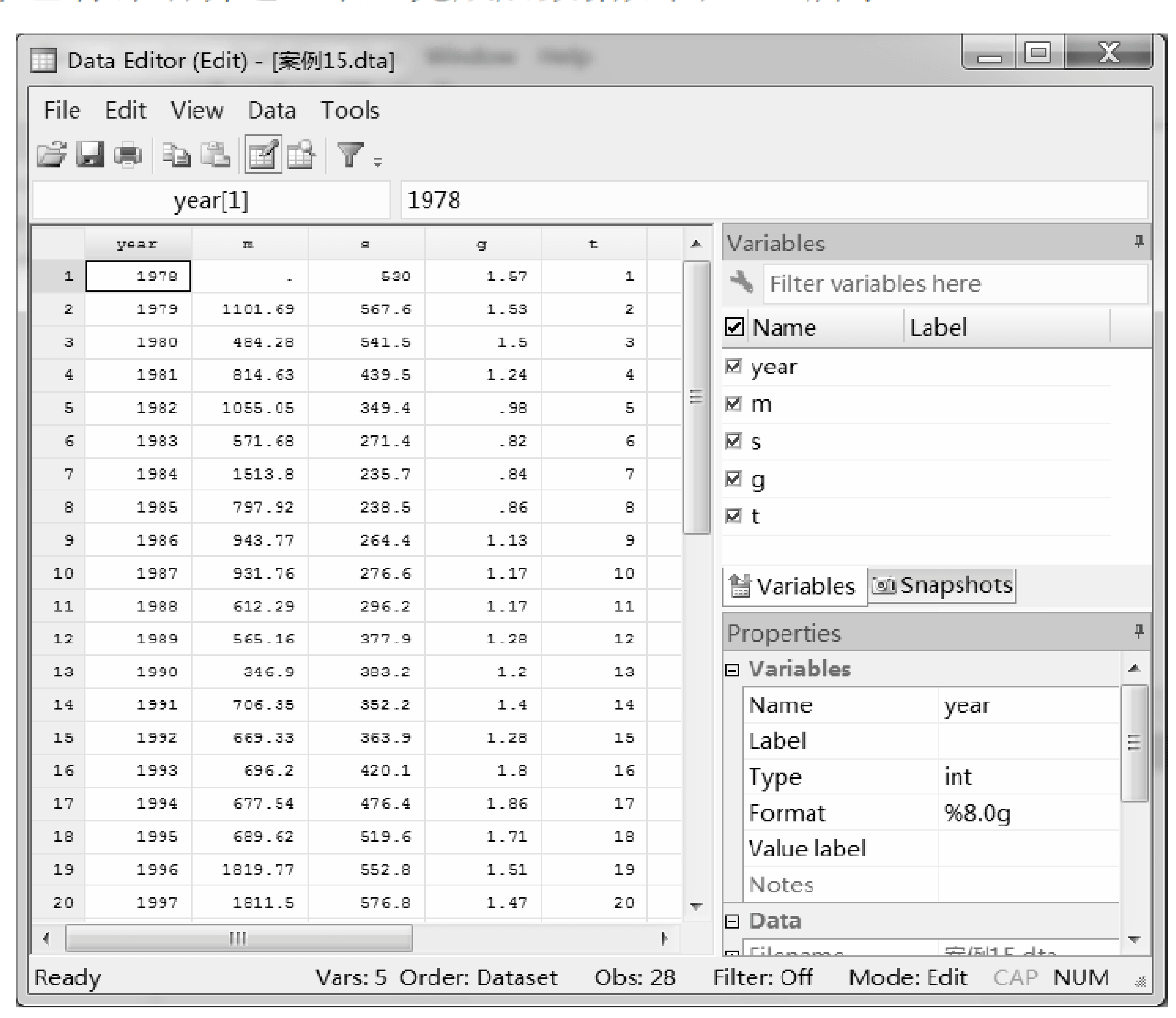
【例 15.1】农村家庭联产承包责任制的推行，以及城市化进程的加快，使得我国大批劳动力从农村解放出来，向当地乡镇企业和城市转移。农村劳动力的大批转移，有效改善了我国劳动力的整体利用状况，提高了人力资源的市场配置效率，对农村经济乃至整个国民经济的发展都起到了非常大的推动作用。那么影响农村劳动力转移的因素有哪些呢？某课题组对该问题进行了实证研究。该课题组选择的具有代表性的变量和数据如表 15.1 所示。试将数据整理成 Stata 数据文件，并进行简要分析。

表 15.1　农村人口城乡转移规模年度数据及相关变量数据

年份	城乡人口净转移/万人	城镇失业规模/万人	城乡收入差距	制度因素
1978		530	1.57	1
1979	1101.69	567.6	1.53	2
1980	484.28	541.5	1.5	3
1981	814.63	439.5	1.24	4
1982	1055.05	349.4	0.98	5
1983	571.68	271.4	0.82	6
…	…	…	…	…
2001	1832.07	681	1.9	24
2002	1814.92	770	2.11	25
2003	1821.55	800	2.23	26
2004	1779.12	827	2.21	27
2005	1785.18	839	2.22	28

15.1.3　Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有 5 个变量，分别为年份、城乡人口净转移、城镇失业规模、城乡收入差距和制度因素。我们把年份变量设定为 year，把城乡人口净转移变量设定为 m，把城镇失业规模变量设定为 s，把城乡收入差距变量设定为 g，把制度因素变量设定为 t，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作在第 1 章中已有详细讲述。录入完成后数据如图 15.1 所示。



	year	m	s	g	t
1	1978	.	530	1.57	1
2	1979	1101.69	567.6	1.53	2
3	1980	484.28	541.5	1.5	3
4	1981	814.63	439.5	1.24	4
5	1982	1055.05	349.4	.98	5
6	1983	571.68	271.4	.82	6
7	1984	1513.8	235.7	.84	7
8	1985	797.92	238.5	.86	8
9	1986	943.77	264.4	1.13	9
10	1987	931.76	276.6	1.17	10
11	1988	612.29	296.2	1.17	11
12	1989	565.16	377.9	1.28	12
13	1990	346.9	383.2	1.2	13
14	1991	706.35	352.2	1.4	14
15	1992	669.33	363.9	1.28	15
16	1993	696.2	420.1	1.8	16
17	1994	677.54	476.4	1.86	17
18	1995	689.62	519.6	1.71	18
19	1996	1819.77	552.8	1.51	19
20	1997	1811.5	576.8	1.47	20

图 15.1　案例 15.1 数据

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中分别输入如下命令并按键盘上的回车键进行确认。

- regress m s g t：本命令的含义是不考虑数据的时间序列性质，直接以城乡人口净转移变量为因变量，以城镇失业规模、城乡收入差距、制度因素为自变量，对数据进行多重线性回归。
- tsset year：本命令的含义是把年份作为日期变量对数据进行时间序列定义。
- twoway(line m year)：本命令的含义是绘制时间序列趋势图来描述变量城乡人口净转移随时间的变动趋势。
- twoway(line s year)：本命令的含义是绘制时间序列趋势图来描述变量城镇失业规模随时间的变动趋势。
- twoway(line g year)：本命令的含义是绘制时间序列趋势图来描述变量城乡收入差距随时间的变动趋势。
- twoway(line t year)：本命令的含义是绘制时间序列趋势图来描述变量制度因素随时间的变动趋势。
- twoway(line d.m year)：本命令的含义是绘制时间序列趋势图来描述变量城乡人口净转移的一阶差分随时间的变动趋势。
- twoway(line d.s year)：本命令的含义是绘制时间序列趋势图来描述变量城镇失业规模的一阶差分随时间的变动趋势。
- twoway(line d.g year)：本命令的含义是绘制时间序列趋势图来描述变量城乡收入差距的一阶差分随时间的变动趋势。
- twoway(line d.t year)：本命令的含义是绘制时间序列趋势图来描述变量制度因素的一阶差分随时间的变动趋势。

03 设置完毕后，等待输出结果。

15.1.4 结果分析

在 Stata 14.0 主界面的结果窗口可以看到如图 15.2~图 15.11 所示的分析结果。

分析结果 1 是不考虑数据的时间序列性质，直接对数据进行简单回归的结果。

```
. regress  m s g t

      Source |       SS           df       MS      Number of obs   =        27
-------------+----------------------------------   F(  3,     23)  =     17.50
       Model |  5572311.68         3  1857437.23   Prob > F        =    0.0000
    Residual |  2441241.24        23  106140.923   R-squared       =    0.6954
-------------+----------------------------------   Adj R-squared   =    0.6556
       Total |  8013552.92        26  308213.574   Root MSE        =    325.79

------------------------------------------------------------------------------
           m |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           s |   3.498603   .8786972     3.98   0.001     1.680879    5.316327
           g |  -1408.282   422.5061    -3.33   0.003    -2282.303   -534.2617
           t |    47.3141   13.75179     3.44   0.002     18.86635    75.76185
       _cons |   850.7036   272.2616     3.12   0.005     287.4877     1413.92
------------------------------------------------------------------------------
```

图 15.2　分析结果 1

从上述分析结果中可以看出共有 27 个样本参与了分析，模型的 F 值(3, 23) = 17.50，P 值

（Prob > F）＝0.0000，说明模型整体上是非常显著的。模型的可决系数（R-squared）为 0.6954，模型修正的可决系数（Adj R-squared）为 0.6556，说明模型的解释能力还是差强人意的。

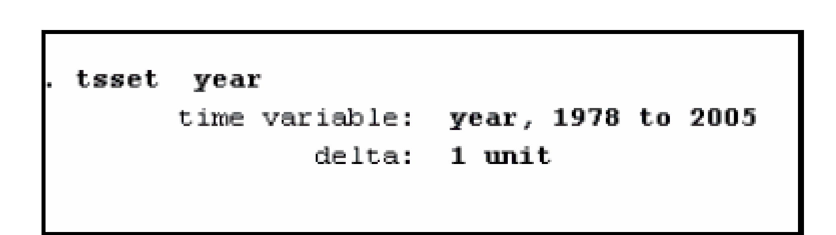
模型的回归方程是：

m=3.498603* s -1408.282*g+47.3141*t+850.7036

变量 s 的系数标准误是 0.8786972，t 值为 3.98，P 值为 0.001，系数是非常显著的，95%的置信区间为[1.680879,5.316327]。变量 g 的系数标准误是 422.5061，t 值为-3.33，P 值为 0.003，系数也是非常显著的，95%的置信区间为[-2282.303, -534.2617]。变量 t 的系数标准误是 13.75179，t 值为 3.44，P 值为 0.002，系数也是非常显著的，95%的置信区间为[18.86635, 75.76185]。常数项的系数标准误是 272.2616，t 值为 3.12，P 值为 0.005，系数也是非常显著的，95%的置信区间为[287.4877 , 1413.92]。

从上面的分析可以看出简单回归的模型在一定程度上是可以接受的，但也存在提升改进的空间。本模型得到的基本结论是城乡人口转移规模（m）随着城乡实际收入差距（g）的扩大而扩大；城镇失业规模（s）对农村劳动力转移具有阻碍作用；制度因素（t）对农村劳动力转移的制约作用逐渐下降。

分析结果 2 显示的是我们把年份作为日期变量对数据进行时间定义的结果，如图 15.3 所示。



```
. tsset  year
        time variable:  year, 1978 to 2005
                delta:  1 unit
```

图 15.3　分析结果 2

从上述分析结果中可以看到时间变量是年份（year），区间范围是从 1978 年到 2005 年，间距为 1。

分析结果 3 显示的是变量城乡人口净转移随时间的变动趋势，如图 15.4 所示。

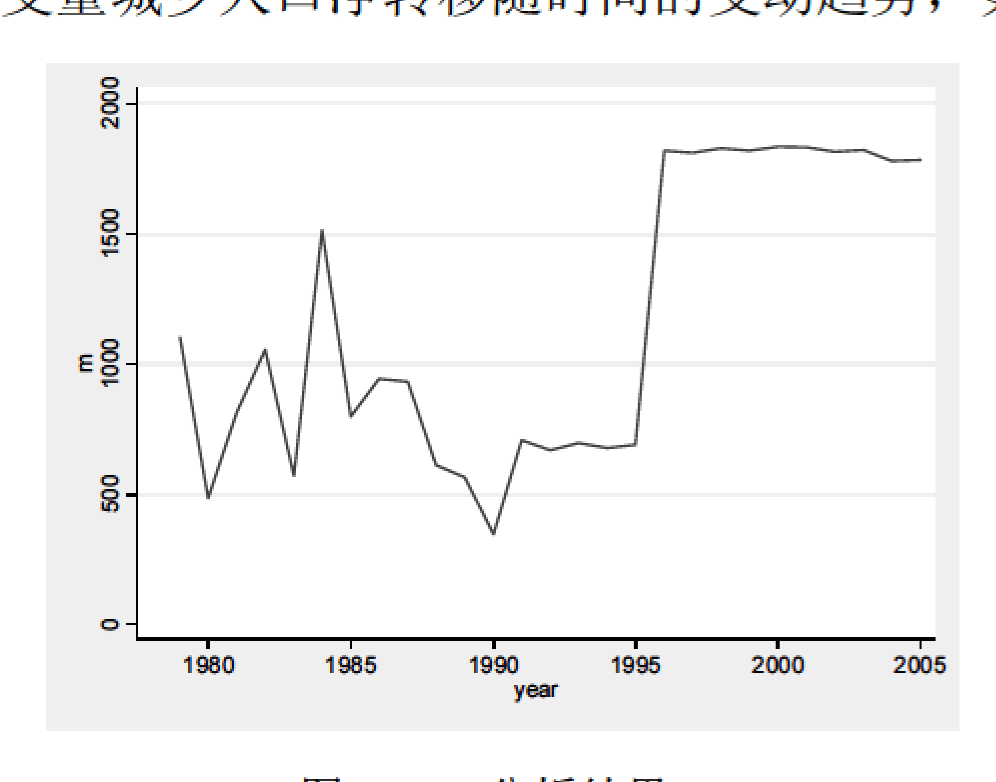


图 15.4　分析结果 3

从上述分析结果中可以看到变量城乡人口净转移没有明显、稳定的长期变化方向。

分析结果 4 显示的是变量城镇失业规模随时间的变动趋势，如图 15.5 所示。

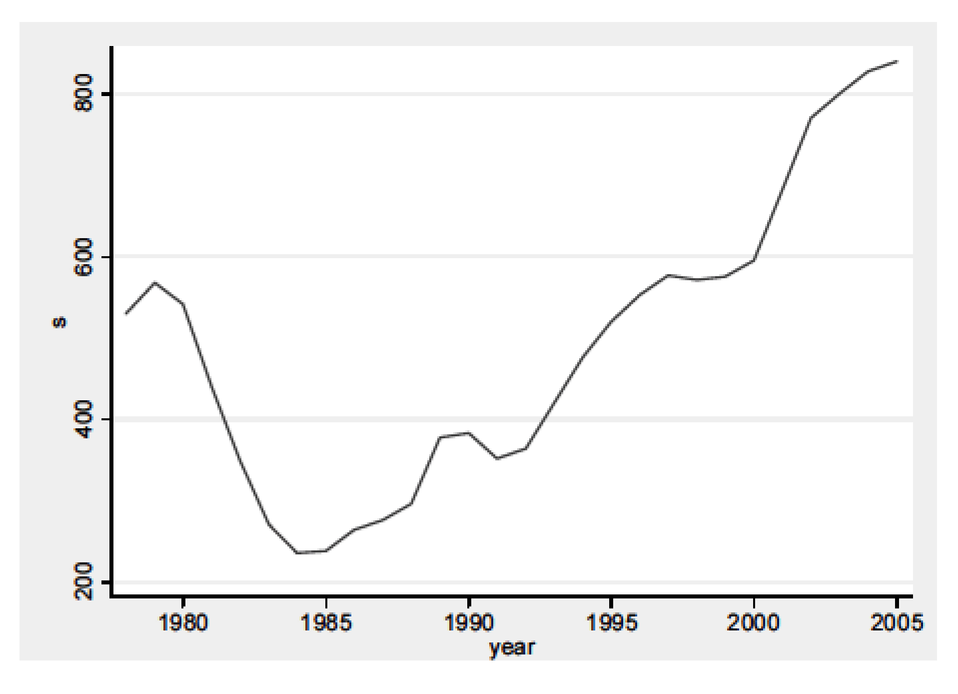


图 15.5　分析结果 4

从上述分析结果中可以看到变量城镇失业规模具有明显、稳定的向上增长趋势。

分析结果 5 显示的是变量城乡收入差距随时间的变动趋势，如图 15.6 所示。

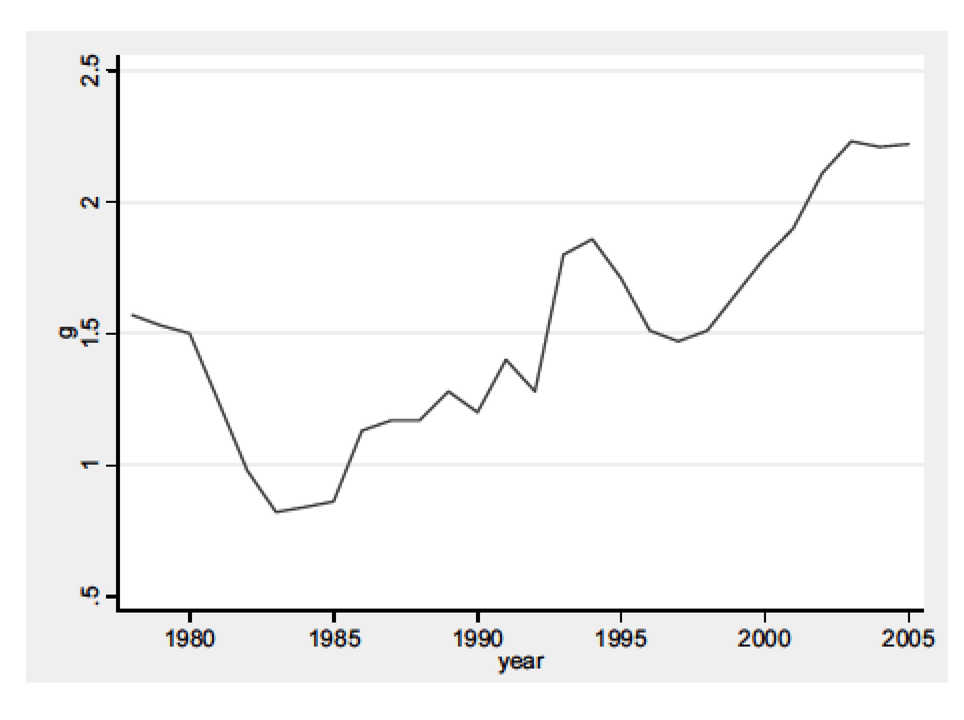


图 15.6　分析结果 5

从上述分析结果中可以看到变量城乡收入差距具有明显、稳定的向上增长趋势。

分析结果 6 显示的是变量制度因素随时间的变动趋势，如图 15.7 所示。

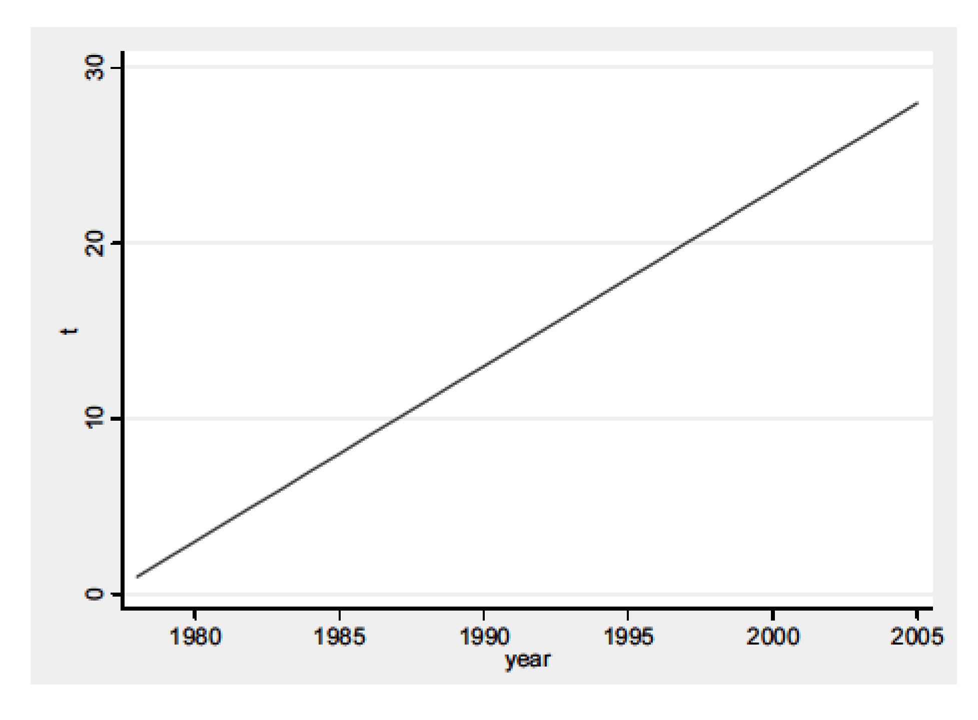


图 15.7　分析结果图 6

从上述分析结果中可以看到变量制度因素具有明显、稳定的向上增长趋势。这是显而易见的。

分析结果 7 显示的是变量城乡人口净转移的增量随时间的变动趋势，如图 15.8 所示。

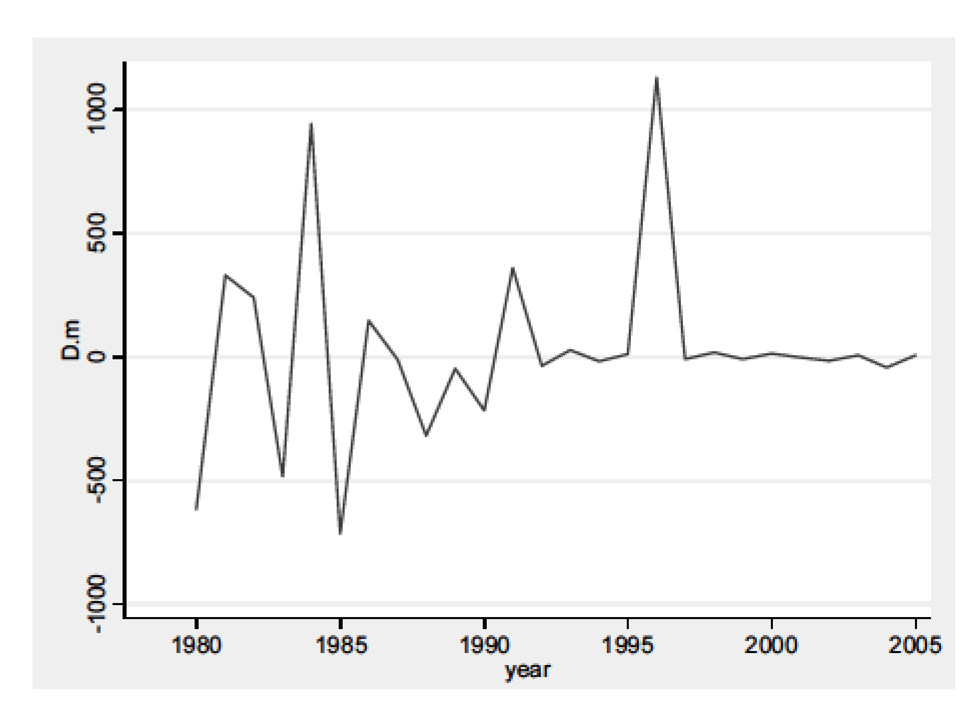


图 15.8 分析结果 7

从上述分析结果中可以看到变量城乡人口净转移的增量没有明显、稳定的长期变化方向。

分析结果 8 显示的是变量城镇失业规模随时间的变动趋势，如图 15.9 所示。

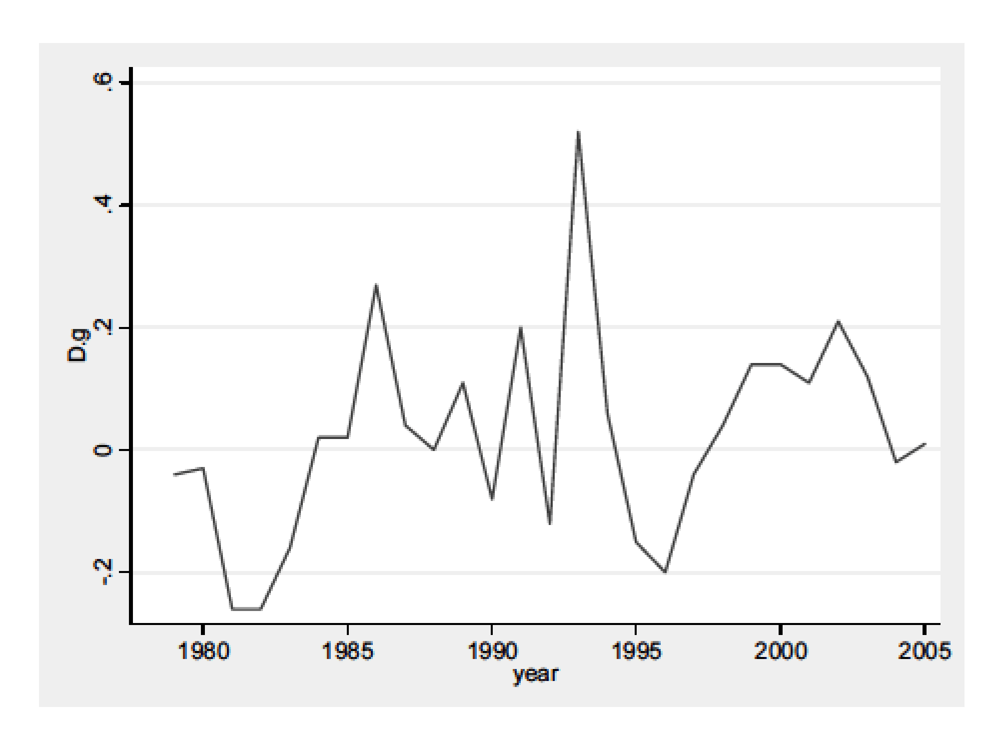


图 15.9 分析结果 8

从上述分析结果中可以看到变量城镇失业规模的增量没有明显、稳定的长期变化方向。

分析结果 9 显示的是变量城乡收入差距随时间的变动趋势，如图 15.10 所示。

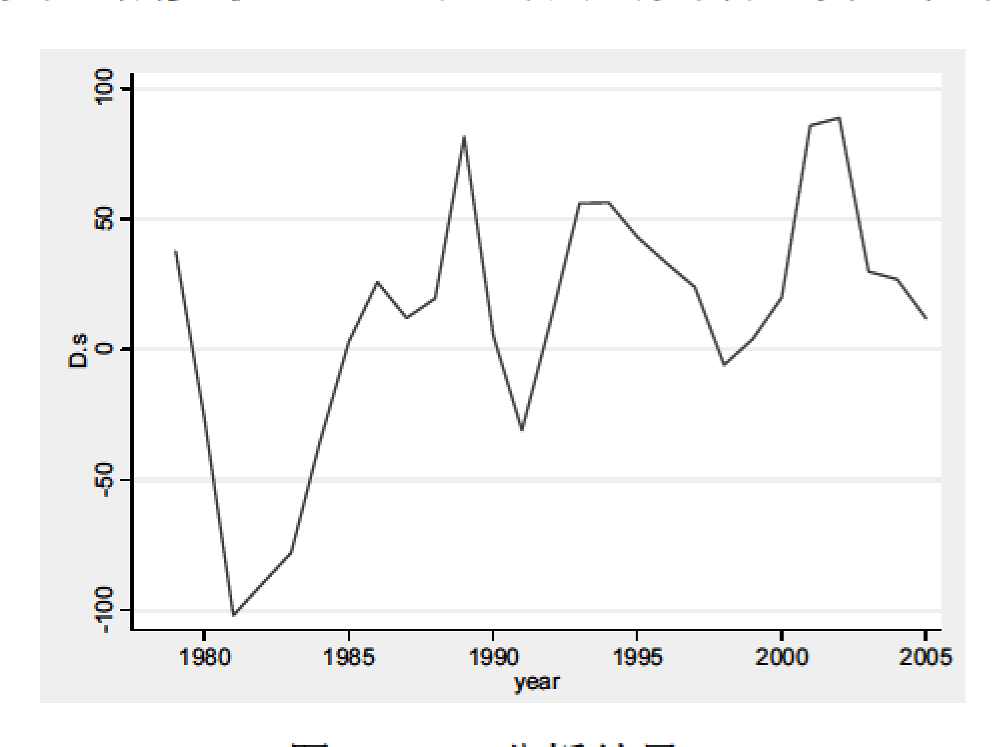


图 15.10 分析结果 9

从上述分析结果中可以看到变量城乡收入差距的增量没有明显、稳定的长期变化方向。

分析结果 10 显示的是变量制度因素的增量随时间的变动趋势，如图 15.11 所示。

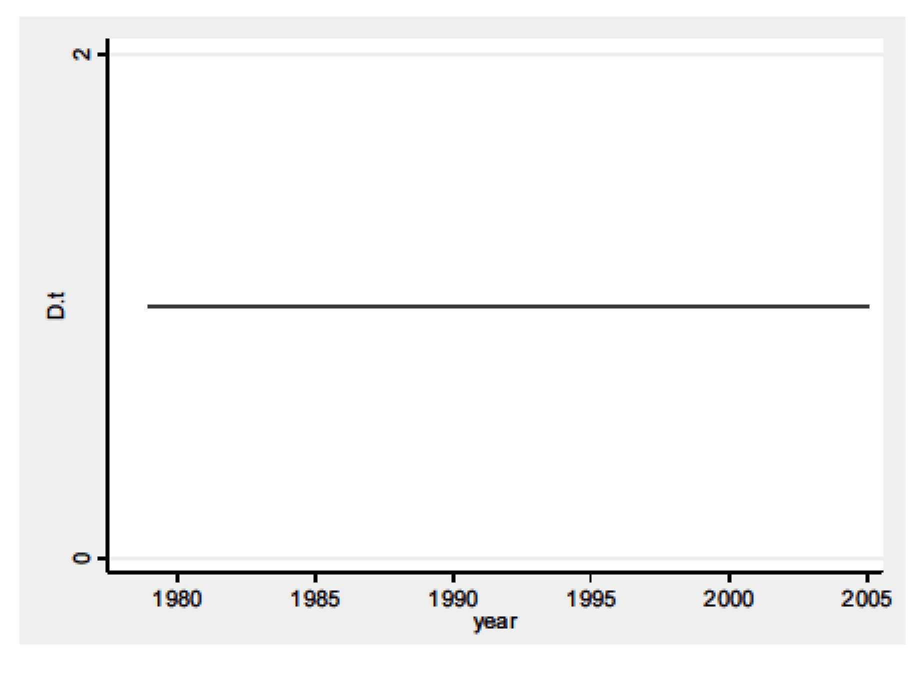


图 15.11　分析结果 10

从上述分析结果中可以看到变量制度因素的增量没有明显、稳定的长期变化方向。

15.1.5　案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

1．延伸 1：清除数据的时间序列格式

例如，我们要把数据恢复为普通的数据，那么操作命令就是：

```
tsset,clear
```

在命令窗口输入命令并按回车键进行确认即可。

2．延伸 2：关于数据处理的一般说明

一般情况下，我们要消除变量的时间序列长期走势后或者说变量平稳后才能进行回归得出有效的结论，所以在绘制变量序列图的时候，如果该变量存在趋势，就应该进行一阶差分后再进行查看。所谓变量的一阶差分指的是对变量的原始数据进行处理，用前面的数据减去后面的数据后得出的一个新的时间序列。如果变量的一阶差分还是存在趋势，就应该进行二阶差分后再进行查看，依次类推，直到数据平稳。所谓二阶差分指的是在把一阶差分得到的时间序列数据作为原始数据，并进行前项减后项处理后得出新的时间序列。一般情况下，如果数据的低阶差分是平稳的，那么高阶差分也是平稳的。

3．延伸 3：关于时间序列运算的有关说明

在上面的案例中，使用了 d.m、d.s、d.g、d.t 等符号分别用来表示 m、s、g、t 等变量的一阶差分。其实还有其他很多简便的运算可供用户使用。常用的 Stata 命令符号与对应的时间序列运算含义如表 15.2 所示。

表 15.2　常用的 Stata 命令符号与对应的时间序列运算含义

Stata命令符号	时间序列运算含义
L.	变量的滞后一期值（Y_{t-1}）
L2.	变量的滞后二期值（Y_{t-2}）
L（1/3）.	变量的滞后一期值到滞后三期值（Y_{t-1} 、Y_{t-2}、Y_{t-3}）

（续表）

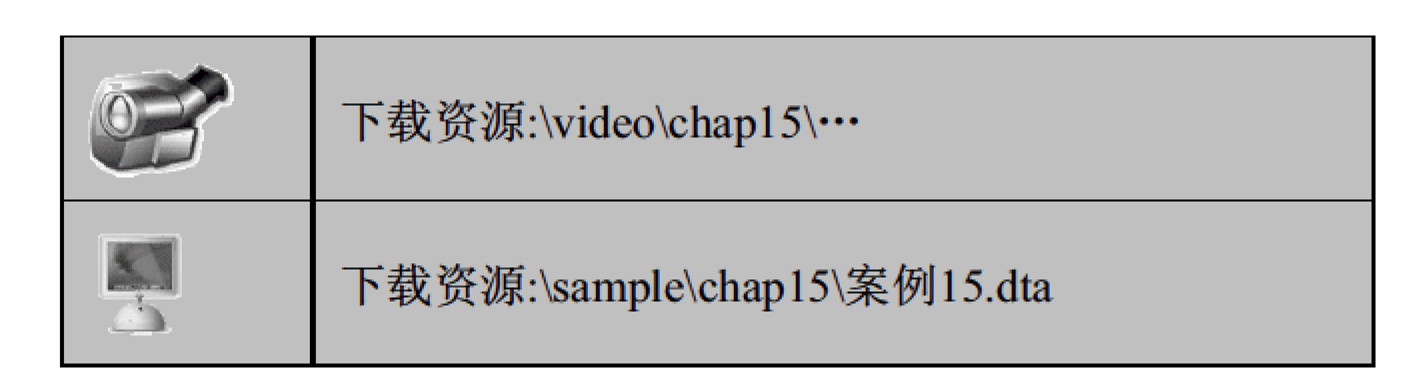
Stata命令符号	时间序列运算含义
F.	变量的向前一期值（Y_{t+1}）
F2.	变量的向前二期值（Y_{t+2}）
D.	变量的一阶差分（Y_t-Y_{t-1}）
D2.	变量的二阶差分（Y_t-Y_{t-1}）-（Y_{t-1}-Y_{t-2}）
S.	变量的季节差分（Y_t-Y_{t-1}），与D.定义相同
S2.	变量的二期季节差分（Y_t-Y_{t-2}），注意与D2.不同

15.2　单位根检验

15.2.1　单位根检验的功能与意义

对于一个时间序列数据而言，数据的平稳性对于模型的构建是非常重要的。如果时间序列数据是不平稳的，可能会导致自回归系数的估计值向左偏向于0，使传统的T检验失效，也有可能会使得两个相互独立的变量出现假相关关系或者回归关系，造成模型结果的失真。在时间序列数据不平稳的情况下，目前公认的能够有效解决假相关或者假回归，构建出合理模型的方法有两种：一种是先对变量进行差分直到数据平稳，再把得到的数据进行回归的方式；另一种就是进行协整检验并构建合理模型的处理方式。那么如何判断数据是否平稳呢？上节中提到的绘制时间序列图的方法可以作为初步推测或者辅助检验的一种方式。但一种更精确的检验方式是：如果数据没有单位根，我们就认为它是平稳的，这时就需要用到本节介绍的单位根检验。

15.2.2　相关数据来源



	下载资源:\video\chap15\…
	下载资源:\sample\chap15\案例15.dta

【例 15.2】本节沿用上节的案例，试通过单位根检验的方式来判断相关变量包括城乡人口净转移、城镇失业规模、城乡收入差距等变量是否平稳。

15.2.3　Stata 分析过程

单位根检验的方式有很多种，此处我们主要介绍常用的两种方式，包括 ADF 检验和 PP 检验。在上一节中，我们通过绘制时间序列趋势图发现城乡人口净转移、城乡人口净转移的一阶差分、城镇失业规模的一阶差分、城乡收入差距的一阶差分是没有时间趋势的，而城镇失业规模和城乡收入差距是有时间趋势的。这些结论将会在后续的操作命令中被用到。

1．ADF 检验

操作步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。
02 在主界面的“Command”文本框中分别输入如下命令并按键盘上的回车键进行确认。

- dfuller m,notrend：本命令的含义是使用 ADF 检验方法，对变量 m 进行单位根检验，不包含时间趋势。
- dfuller s, trend：本命令的含义是使用 ADF 检验方法，对变量 s 进行单位根检验，包含时间趋势。
- dfuller g, trend：本命令的含义是使用 ADF 检验方法，对变量 g 进行单位根检验，包含时间趋势。
- dfuller d.m,notrend：本命令的含义是使用 ADF 检验方法，对变量 d.m 进行单位根检验，不包含时间趋势。
- dfuller d.s, notrend：本命令的含义是使用 ADF 检验方法，对变量 d.s 进行单位根检验，不包含时间趋势。
- dfuller d.g, notrend：本命令的含义是使用 ADF 检验方法，对变量 d.g 进行单位根检验，不包含时间趋势。
- dfuller d2.s, notrend：本命令的含义是使用 ADF 检验方法，对变量 d2.s 进行单位根检验，不包含时间趋势。

03 设置完毕后，等待输出结果。

2．PP 检验

操作步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。
02 在主界面的“Command”文本框中分别输入如下命令并按键盘上的回车键进行确认。

- pperron m,notrend：本命令的含义是使用 PP 检验方法，对变量 m 进行单位根检验，不包含时间趋势。
- pperron s, trend：本命令的含义是使用 PP 检验方法，对变量 s 进行单位根检验，包含时间趋势。
- pperron g, trend：本命令的含义是使用 PP 检验方法，对变量 g 进行单位根检验，包含时间趋势。
- pperron d.m,notrend：本命令的含义是使用 PP 检验方法，对变量 d.m 进行单位根检验，不包含时间趋势。
- pperron d.s, notrend：本命令的含义是使用 PP 检验方法，对变量 d.s 进行单位根检验，不包含时间趋势。
- pperron d.g, notrend：本命令的含义是使用 PP 检验方法，对变量 d.g 进行单位根检验，不包含时间趋势。
- pperron d2.s, notrend：本命令的含义是使用 PP 检验方法，对变量 d2.s 进行单位根检

验，不包含时间趋势。

03 设置完毕后，等待输出结果。

15.2.4 结果分析

在 Stata 14.0 主界面的结果窗口可以看到如图 15.12~图 15.25 所示的分析结果。

1. ADF 检验结果

ADF 检验的结果如图 15.12~图 15.18 所示。其中，图 15.12 展示的是城乡人口净转移这一变量的 ADF 检验结果。

```
. dfuller m,notrend

Dickey-Fuller test for unit root                   Number of obs   =        26

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -1.617            -3.743            -2.997            -2.629
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.4745
```

图 15.12 城乡人口净转移

ADF 检验的原假设是数据有单位根。从上面的结果中可以看出 P 值（MacKinnon approximate p-value for Z(t)）为 0.4745，接受了有单位根的原假设，这一点也可以通过观察 Z(t)值得到。实际 Z(t)值为-1.617，在 1%的置信水平（-3.743）、5%的置信水平（-2.997）、10%的置信水平上（-2.629）都无法拒绝原假设，所以城乡人口净转移这一变量数据是存在单位根的，需要对其做一阶差分后再继续进行检验。

图 15.13 展示的是城镇失业规模这一变量的 ADF 检验结果。

```
. dfuller s, trend

Dickey-Fuller test for unit root                   Number of obs   =        27

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -1.821            -4.362            -3.592            -3.235
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.6948
```

图 15.13 城镇失业规模

ADF 检验的原假设是数据有单位根。从上面的结果中可以看出 P 值（MacKinnon approximate p-value for Z(t)）为 0.6948，接受了有单位根的原假设，这一点也可以通过观察 Z(t)值得到。实际 Z(t)值为-1.821，在 1%的置信水平（-4.362）、5%的置信水平（-3.592）、10%的置信水平上（-3.235）都无法拒绝原假设，所以城镇失业规模这一变量数据是存在单位根的，需要对其做一阶差分后再继续进行检验。

图 15.14 展示的是城乡收入差距这一变量的 ADF 检验结果。

```
. dfuller g, trend

Dickey-Fuller test for unit root                   Number of obs   =        27

                          ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -2.435            -4.362            -3.592            -3.235
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.3612
```

图 15.14　城乡收入差距

ADF 检验的原假设是数据有单位根。从上面的结果中可以看出 P 值（MacKinnon approximate p-value for Z(t)）为 0.3612，接受了有单位根的原假设，这一点也可以通过观察 Z(t)值得到。实际 Z(t)值为-2.435，在 1%的置信水平（-4.362）、5%的置信水平（-3.592）、10%的置信水平上（-3.235）都无法拒绝原假设，所以城乡收入差距这一变量数据是存在单位根的，需要对其做一阶差分再继续进行检验。

图 15.15 展示的是城乡人口净转移这一变量的一阶差分的 ADF 检验结果。

```
. dfuller d.m,notrend

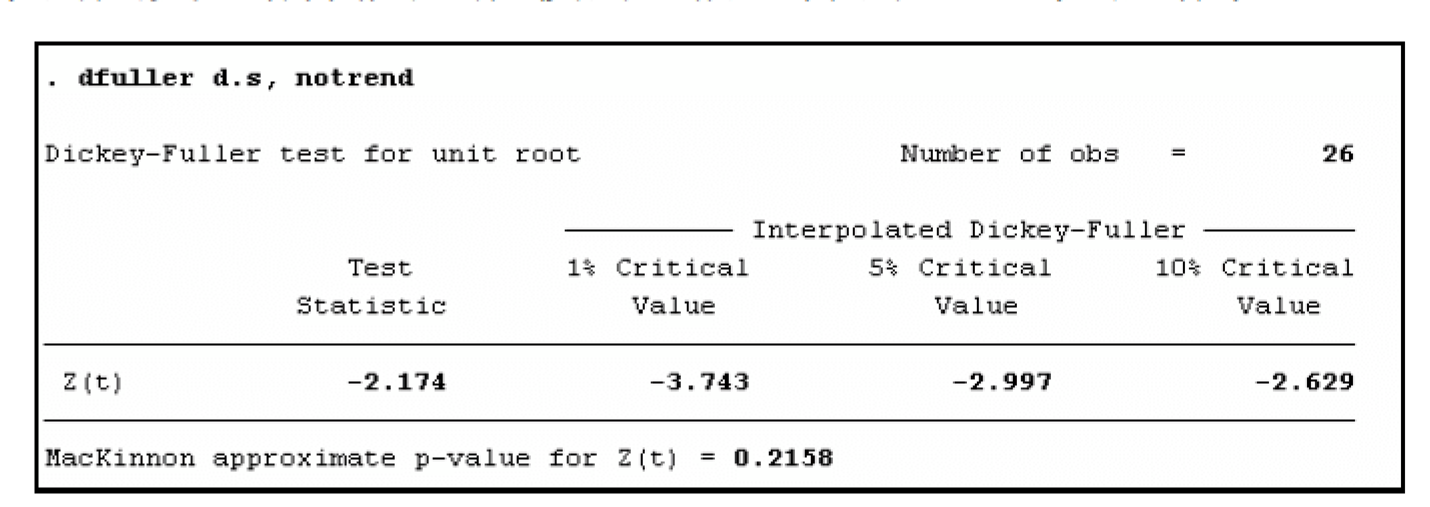
Dickey-Fuller test for unit root                   Number of obs   =        25

                          ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -8.085            -3.750            -3.000            -2.630
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0000
```

图 15.15　城乡人口净转移一阶差分

ADF 检验的原假设是数据有单位根。从上面的结果中可以看出 P 值（MacKinnon approximate p-value for Z(t)）为 0.0000，拒绝了有单位根的原假设，这一点也可以通过观察 Z(t)值得到。实际 Z(t)值为-8.085，在 1%的置信水平（-3.750）、5%的置信水平（-3.000）、10%的置信水平上（-2.630）都应拒绝原假设，所以城乡人口净转移这一变量的一阶差分数据是不存在单位根的。

图 15.16 展示的是变量城镇失业规模的一阶差分的 ADF 检验结果。



```
. dfuller d.s, notrend

Dickey-Fuller test for unit root                   Number of obs   =        26

                          ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -2.174            -3.743            -2.997            -2.629
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.2158
```

图 15.16　镇失业规模一阶差分

ADF 检验的原假设是数据有单位根。从上面的结果中可以看出 P 值（MacKinnon approximate p-value for Z(t)）为 0.2158，接受了有单位根的原假设，这一点也可以通过观察 Z(t)值得到。实际 Z(t)值为-2.174，在 1%的置信水平（-3.743）、5%的置信水平（-2.997）、10%的置信水平上（-2.629）都无法拒绝原假设，所以城镇失业规模这一变量的一阶差分数据是存在单位根的，需要对城镇失业规模做二阶差分后再继续进行检验。

图 15.17 展示的是变量城乡收入差距的一阶差分的 ADF 检验结果。

```
. dfuller d.g, notrend

Dickey-Fuller test for unit root                   Number of obs   =        26

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -4.016            -3.743            -2.997            -2.629
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0013
```

图 15.17　城乡收入差距一阶差分

ADF 检验的原假设是数据有单位根。从上面的结果中可以看出 P 值（MacKinnon approximate p-value for Z(t)）为 0.0013，拒绝了有单位根的原假设，这一点也可以通过观察 Z(t)值得到。实际 Z(t)值为-4.016，在 1%的置信水平（-3.743）、5%的置信水平（-2.997）、10%的置信水平上（-2.629）都拒绝原假设，所以城乡收入差距这一变量的一阶差分数据是不存在单位根的。

图 15.18 展示的是变量城镇失业规模的二阶差分的 ADF 检验结果。

```
. dfuller d2.s, notrend

Dickey-Fuller test for unit root                   Number of obs   =        25

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -4.192            -3.750            -3.000            -2.630
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0007
```

图 15.18　城镇失业规模的二阶差分

ADF 检验的原假设是数据有单位根。从上面的结果中可以看出 P 值（MacKinnon approximate p-value for Z(t)）为 0.0007，拒绝了有单位根的原假设。这一点也可以通过观察 Z(t)值得到。实际 Z(t)值为-4.192，在 1%的置信水平（-3.750）、5%的置信水平（-3.000）、10%的置信水平上（-2.630）都拒绝原假设，所以城镇失业规模这一变量的二阶差分数据是不存在单位根的。

2. PP 检验结果

PP 检验的结果如图 15.19~图 15.25 所示。其中，图 15.19 展示的是城乡人口净转移这一变量的 PP 检验结果。

```
. pperron m,notrend

Phillips-Perron test for unit root                 Number of obs   =        26
                                                   Newey-West lags =         2

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(rho)           -4.460           -17.268           -12.532           -10.220
 Z(t)             -1.409            -3.743            -2.997            -2.629
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.5779
```

图 15.19　城乡人口净转移

PP 检验的原假设是数据有单位根。从上面的结果中可以看出 P 值（MacKinnon approximate p-value for Z(t)）为 0.5779，接受了有单位根的原假设，这一点也可以通过观察 Z(t)值和 Z(rho)值得到。实际 Z(t)值为-1.409，在 1%的置信水平（-3.743）、5%的置信水平（-2.997）、10%的置信水平上（-2.629）都无法拒绝原假设。实际 Z(rho)值为-4.460，在 1%的置信水平（-17.268）、5%的置信水平（-12.532）、10%的置信水平上（-10.220）都无法拒绝原假设，所以城乡人口净转移这一变量数据是存在单位根的，需要对其做一阶差分后再继续进行检验。

图 15.20 展示的是城镇失业规模这一变量的 PP 检验结果。

```
. pperron s, trend

Phillips-Perron test for unit root                   Number of obs   =        27
                                                     Newey-West lags =         2

                                   ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(rho)          -3.426          -22.756           -18.052           -15.696
 Z(t)            -1.800           -4.362            -3.592            -3.235
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.7048
```

图 15.20　城镇失业规模

PP 检验的原假设是数据有单位根。从上面的结果中可以看出 P 值（MacKinnon approximate p-value for Z(t)）为 0.7048，接受了有单位根的原假设，这一点也可以通过观察 Z(t)值和 Z(rho)值得到。实际 Z(t)值为-1.800，在 1%的置信水平（-4.362）、5%的置信水平（-3.592）、10%的置信水平上（-3.235）都无法拒绝原假设。实际 Z(rho)值为-3.426，在 1%的置信水平（-22.756）、5%的置信水平（-18.052）、10%的置信水平上（-15.696）都无法拒绝原假设，所以城镇失业规模这一变量数据是存在单位根的，需要对其做一阶差分后再继续进行检验。

图 15.21 展示的是城乡收入差距这一变量的 PP 检验结果。

```
. pperron g, trend

Phillips-Perron test for unit root                   Number of obs   =        27
                                                     Newey-West lags =         2

                                   ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(rho)          -7.547          -22.756           -18.052           -15.696
 Z(t)            -2.459           -4.362            -3.592            -3.235
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.3489
```

图 15.21　城乡收入差距

PP 检验的原假设是数据有单位根。从上面的结果中可以看出 P 值（MacKinnon approximate p-value for Z(t)）为 0.3489，接受了有单位根的原假设，这一点也可以通过观察 Z(t)值和 Z(rho)值得到。实际 Z(t)值为-2.459，在 1%的置信水平（-4.362）、5%的置信水平（-3.592）、10%的置信水平上（-3.235）都无法拒绝原假设。实际 Z(rho)值为-7.547，在 1%的置信水平（-22.756）、5%的置信水平（-18.052）、10%的置信水平上（-15.696）都无法拒绝原假设，所以城乡收入差距这一变量数据是存在单位根的，需要对其做一阶差分后再继续进行检验。

图 15.22 展示的是城乡人口净转移这一变量的一阶差分的 PP 检验结果。

```
. pperron d.m,notrend

Phillips-Perron test for unit root                   Number of obs   =        25
                                                     Newey-West lags =         2

                               ———————— Interpolated Dickey-Fuller ————————
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(rho)          -35.522          -17.200           -12.500           -10.200
 Z(t)             -8.079           -3.750            -3.000            -2.630
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0000
```

图 15.22　城乡人口净转移一阶差分

PP 检验的原假设是数据有单位根。从上面的结果中可以看出 P 值（MacKinnon approximate p-value for Z(t)）为 0.0000，拒绝了有单位根的原假设，这一点也可以通过观察 Z(t)值和 Z(rho)值得到。实际 Z(t)值为-8.079，在 1%的置信水平（-3.750）、5%的置信水平（-3.000）、10%的置信水平上（-2.630）都应拒绝原假设。实际 Z(rho)值为-35.522，在 1%的置信水平（-17.200）、5%的置信水平（-12.500）、10%的置信水平上（-10.200）都应拒绝原假设，所以城乡人口净转移这一变量的一阶差分数据是不存在单位根的。

图 15.23 展示的是变量城镇失业规模的一阶差分的 PP 检验结果。

```
. pperron d.s, notrend

Phillips-Perron test for unit root                   Number of obs   =        26
                                                     Newey-West lags =         2

                               ———————— Interpolated Dickey-Fuller ————————
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(rho)          -10.379          -17.268           -12.532           -10.220
 Z(t)             -2.386           -3.743            -2.997            -2.629
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.1457
```

图 15.23　城镇失业规模一阶差分

PP 检验的原假设是数据有单位根。从上面的结果中可以看出 P 值（MacKinnon approximate p-value for Z(t)）为 0.1457，接受了有单位根的原假设，这一点也可以通过观察 Z(t)值和 Z(rho)值得到。实际 Z(t)值为-2.386，在 1%的置信水平（-3.743）、5%的置信水平（-2.997）、10%的置信水平上（-2.629）都无法拒绝原假设。实际 Z(rho)值为-10.379，在 1%的置信水平（-17.268）、5%的置信水平（-12.532）、10%的置信水平上（-10.220）都无法拒绝原假设，所以城镇失业规模这一变量的一阶差分数据是存在单位根的，需要对城镇失业规模做二阶差分后再继续进行检验。

图 15.24 展示的是变量城乡收入差距的一阶差分的 PP 检验结果。

```
. pperron d.g, notrend

Phillips-Perron test for unit root                   Number of obs   =        26
                                                     Newey-West lags =         2

                               ———————— Interpolated Dickey-Fuller ————————
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(rho)          -21.701          -17.268           -12.532           -10.220
 Z(t)             -4.051           -3.743            -2.997            -2.629
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0012
```

图 15.24　城乡收入差距一阶差分

PP 检验的原假设是数据有单位根。从上面的结果中可以看出 P 值（MacKinnon approximate p-value for Z(t)）为 0.0012，拒绝了有单位根的原假设，这一点也可以通过观察 Z(t)值和 Z(rho)值得到。实际 Z(t)值为-4.051，在 1%的置信水平（-3.743）、5%的置信水平（-2.997）、10%的置信水平上（-2.629）都拒绝原假设。实际 Z(rho)值为-21.701，在 1%的置信水平（-17.268）、5%的置信水平（-12.532）、10%的置信水平上（-10.220）都应拒绝原假设，所以城乡收入差距这一变量的一阶差分数据是不存在单位根的。

图 15.25 展示的是变量城镇失业规模的二阶差分的 PP 检验结果。

```
. pperron d2.s, notrend

Phillips-Perron test for unit root                 Number of obs   =        25
                                                   Newey-West lags =         2

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(rho)         -17.168           -17.200           -12.500           -10.200
 Z(t)            -4.176            -3.750            -3.000            -2.630
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0007
```

图 15.25　城镇失业规模二阶差分

PP 检验的原假设是数据有单位根。从上面的结果中可以看出 P 值（MacKinnon approximate p-value for Z(t)）为 0.0007，拒绝了有单位根的原假设。这一点也可以通过观察 Z(t)值和 Z(rho)值得到。实际 Z(t)值为-4.176，在 1%的置信水平（-3.750）、5%的置信水平（-3.000）、10%的置信水平上（-2.630）都拒绝原假设。实际 Z(rho)值为-17.168，在 1%的置信水平（-17.200）、5%的置信水平（-12.500）、10%的置信水平上（-10.200）都应拒绝原假设，所以城镇失业规模这一变量的二阶差分数据是不存在单位根的。

可以看出，在本例中 ADF 检验结果和 PP 检验结果是完全一致的，所以，通过比较可以有把握地认为城乡人口净转移、城乡收入差距两个变量是一阶单整的，而城镇失业规模变量是二阶单整的。

15.2.5　案例延伸

按照前面讲述的解决方法，可以对变量进行相应阶数的差分，然后进行回归，即可避免出现伪回归的情况。

构建如下所示的模型方程：

d.m=a*d.g+b*d2.s+c*t+u

其中，a、b、c 为系数，u 为误差扰动项。

在主界面的“Command”文本框中输入如下命令并按键盘上的回车键进行确认：

```
regress d.m d2.s d.g t
```

即可出现如图 15.26 所示的回归分析结果。

```
. regress d.m d2.s d.g t

      Source |       SS           df       MS      Number of obs   =        26
-------------+----------------------------------   F(3, 22)        =      0.26
       Model |   127232.42         3  42410.8068   Prob > F        =    0.8551
    Residual |  3621825.92        22  164628.451   R-squared       =    0.0339
-------------+----------------------------------   Adj R-squared   =   -0.0978
       Total |  3749058.34        25  149962.334   Root MSE        =    405.74

------------------------------------------------------------------------------
         D.m |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           s |
         D2. |   .8166687   2.190912     0.37   0.713    -3.727005    5.360342
             |
           g |
         D1. |  -374.9964   525.5279    -0.71   0.483    -1464.875    714.8818
             |
           t |   7.656357    11.1856     0.68   0.501    -15.54116    30.85387
       _cons |  -81.62952   187.2142    -0.44   0.667     -469.888    306.6289
------------------------------------------------------------------------------
```

图 15.26　分析结果图

从上述分析结果中可以看到，结果与本章开始在数据无处理状态下进行的“伪回归”的结果是不同的。可以看出共有 26 个样本参与了分析，这是因为进行差分会减少观测样本。模型的 F 值(3, 22) = 0.26，P 值（Prob > F） = 0.8551，说明模型整体上是不显著的，本章开始得出的结果其实是一种真真正正的“伪回归”。模型的可决系数（R-squared）为 0.0339，模型修正的可决系数（Adj R-squared）为-0.0978，说明模型几乎没有什么解释能力。

模型的回归方程是：

d.m=0.8166687* d2.s-374.9964*d1.g+7.656357*t+-81.62952

变量 d2.s 的系数标准误是 2.190912，t 值为 0.37，P 值为 0.713，系数是非常不显著的，95%的置信区间为[-3.727005,5.360342]。变量 d1.g 的系数标准误是 525.5279，t 值为-0.71，P 值为 0.483，系数也是非常显著的，95%的置信区间为[-1464.875, 714.8818]。变量 t 的系数标准误是 11.1856，t 值为 0.68，P 值为 0.501，系数也是非常显著的，95%的置信区间为[-15.54116 , 30.85387]。常数项的系数标准误是 187.2142，t 值为-0.44，P 值为 0.667，系数也是非常显著的，95%的置信区间为[-469.888, 306.6289]。

从上面的分析可以看出，本模型得到的基本结论是城乡人口转移规模（m）随着城乡实际收入差距（g）的扩大而扩大；城镇失业规模（s）对农村劳动力转移具有阻碍作用；制度因素（t）对农村劳动力转移的制约作用逐渐下降，这一点与伪回归得出的结果是一致的。

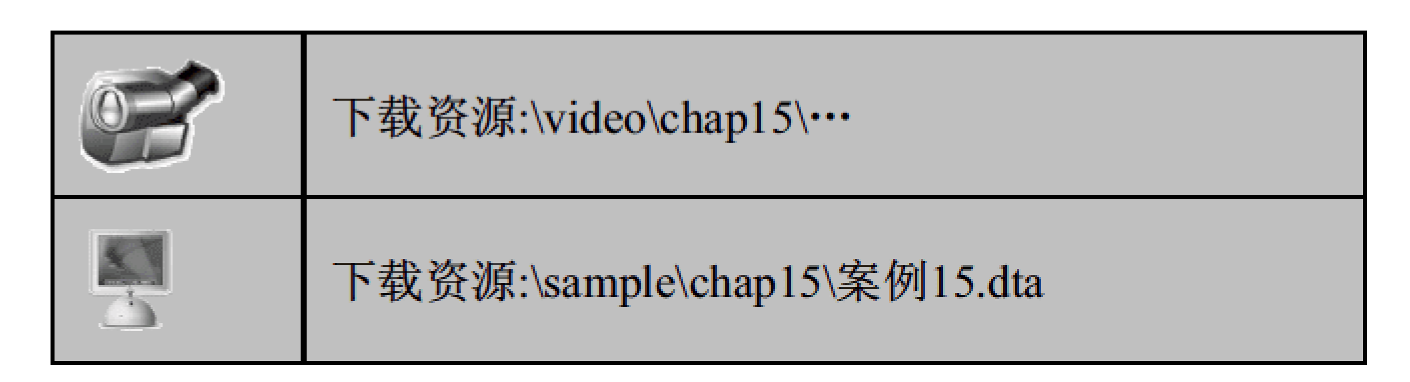
15.3　协整检验

15.3.1　协整检验的功能与意义

在上一节中，我们提到对于一个时间序列数据而言，数据的平稳性对于模型的构建是非常重要的。在时间序列数据不平稳的情况下，构建出合理模型的另外一种方法就是进行协整检验并构建合理模型。协整的思想就是把存在一阶单整的变量放在一起进行分析，通过这些变量进行线性组合，从而消除它们的随机趋势，得到其长期联动趋势。目前学者公认的协整检验的

有效方法有两种：一种是 EG-ADF 检验；另外一种是迹检验。一般认为，迹检验的效果要好于 EG-ADF 检验，但 EG-ADF 作为传统经典的检验方法应用范围要更广一些。下面就来介绍一下协整检验在实例中的应用。

15.3.2 相关数据来源



	下载资源:\video\chap15\…
	下载资源:\sample\chap15\案例15.dta

【例 15.3】本节沿用上节的案例，试通过 EG-ADF 检验、迹检验等两种协整检验的方式来判断相关变量包括城乡人口净转移、城镇失业规模、城乡收入差距等变量是否存在长期协整关系。

15.3.3 Stata 分析过程

在前面两节中，通过绘制时间序列趋势图发现城乡人口净转移、城乡人口净转移的一阶差分、城镇失业规模的一阶差分、城乡收入差距的一阶差分是没有时间趋势的，而城镇失业规模和城乡收入差距是有时间趋势的。通过单位根检验发现城乡人口净转移、城乡收入差距两个变量是一阶单整的，而城镇失业规模变量是二阶单整的。这些结论将会在后续的操作命令中被用到。

1. EG-ADF 检验

操作步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中分别输入如下命令并按键盘上的回车键进行确认。

- regress m d.s g：本命令的含义是把城乡人口净转移作为因变量，把城镇失业规模的一阶差分、城乡收入差距作为自变量，用普通最小二乘估计法进行估计。
- predict e,resid：本命令的含义是得到上步回归产生的残差序列。
- twoway(line e year)：本命令的含义是绘制残差序列的时间趋势图。
- dfuller e,notrend nocon lags(1) regress：本命令的含义是对残差序列进行 ADF 检验，观测其是否为平稳序列，其中不包括时间趋势项，不包括常数项，滞后 1 期。

03 设置完毕后，等待输出结果。

2. 迹检验

操作步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中分别输入如下命令并按键盘上的回车键进行确认。

- varsoc m d.s g：本命令旨在根据信息准则确定变量的滞后阶数。
- vecrank m d.s g,lags(4)：本命令旨在确定协整秩。
- vec m d.s g,lags(4) rank(1)：本命令旨在估计协整模型。

03 设置完毕后，等待输出结果。

15.3.4　结果分析

在 Stata 14.0 主界面的结果窗口可以看到如图 15.27~图 15.32 所示的分析结果。

1．EG-ADF 检验

EG-ADF 的检验过程是：首先把城乡人口净转移作为因变量，把城镇失业规模的一阶差分、城乡收入差距作为自变量，用普通最小二乘估计法进行估计得到残差序列，然后对残差序列进行 ADF 检验，观测其是否为平稳序列，如果残差序列是平稳的，那么变量之间的长期协整关系就存在，如果残差序列是不平稳的，那么变量之间的长期协整关系就不存在。本例中，EG-ADF 检验的结果如图 15.27~图 15.30 所示。其中，图 15.27 展示的是把城乡人口净转移作为因变量，把城镇失业规模的一阶差分、城乡收入差距作为自变量，用普通最小二乘估计法进行估计的结果。

```
. regress  m  d.s  g

      Source |       SS           df       MS      Number of obs   =        27
-------------+----------------------------------   F(  2,     24)  =      5.23
       Model |  2433652.47         2  1216826.24   Prob > F        =    0.0130
    Residual |  5579900.45        24  232495.852   R-squared       =    0.3037
-------------+----------------------------------   Adj R-squared   =    0.2457
       Total |  8013552.92        26  308213.574   Root MSE        =    482.18

------------------------------------------------------------------------------
           m |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           s |
         D1. |  -1.229304   2.374201    -0.52   0.609    -6.129415    3.670806
             |
           g |   793.4284   271.4427     2.92   0.007     233.1982    1353.659
       _cons |  -12.01591   401.9297    -0.03   0.976    -841.5581    817.5263
------------------------------------------------------------------------------
```

图 15.27　用普通最小二乘估计法进行估计

从上述分析结果中可以看到共有 27 个样本参与了分析。模型的 F 值(2, 24) ＝ 5.23，P 值（Prob > F）＝0.0130，说明模型整体上是比较显著的。模型的可决系数（R-squared）为 0.3037，模型修正的可决系数（Adj R-squared）为 0.2457，说明模型的解释能力非常一般。

模型的回归方程是：

m=-1.229304* d1.s+793.4284* g-14.01591

变量 d1.s 的系数标准误是 2.374201，t 值为-0.52，P 值为 0.609，系数是非常不显著的，95%的置信区间为[-6.129415,3.670806]。变量 g 的系数标准误是 271.4427，t 值为 2.92，P 值为 0.007，系数也是非常显著的，95%的置信区间为[233.1982 , 1353.659]。常数项的系数标准误是 401.9297，t 值为-0.03，P 值为 0.976，系数也是非常不显著的，95%的置信区间为[-841.5581 , 817.5263]。

图 15.28 展示的是对模型残差的预测结果。

```
. predict e,resid
(1 missing value generated)
```

图 15.28　模型残差的预测结果

图 15.29 展示的是残差序列的时间走势，可以发现残差序列是没有固定时间趋势的。

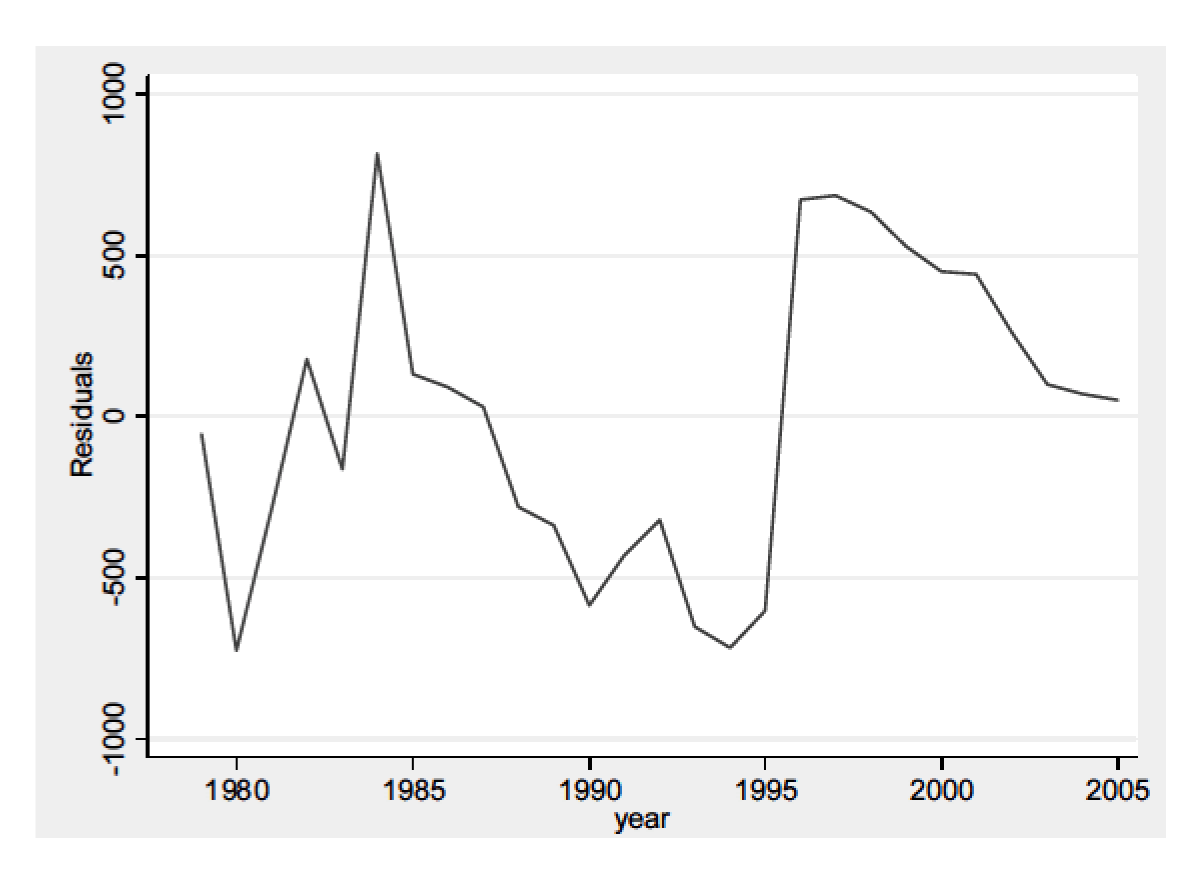


图 15.29　残差序列的时间走势

图 15.30 展示的是残差序列 ADF 检验结果。

```
. dfuller e,notrend nocon lags(1) regress

Augmented Dickey-Fuller test for unit root         Number of obs   =        25

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -2.273            -2.660            -1.950            -1.600

------------------------------------------------------------------------------
         D.e |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           e |
         L1. |  -.3933092   .1730557    -2.27   0.033    -.7513023   -.0353162
         LD. |  -.0295201   .1938465    -0.15   0.880    -.4305222     .371482
------------------------------------------------------------------------------
```

图 15.30　残差序列 ADF 检验结果

ADF 检验的原假设是数据有单位根。从上面的结果中可以看出实际 Z(t)值为-2.273，介于 1%的置信水平（-2.660）和 5%的置信水平（-1.950）之间，所以应该拒绝存在单位根原假设。因此残差序列是不存在单位根的，或者说残差序列是平稳的。

综上所述，城乡人口净转移、城镇失业规模、城乡收入差距 3 个变量存在协整关系。根据上面的分析结果可以构建出相应的模型来描述这种协整关系。关于这一点将在本节的案例延伸部分进行详细说明。

2．迹检验

迹检验的过程是：首先要根据信息准则确定变量的滞后阶数，即模型中变量的个数。信息准则的概念是针对变量的个数，学者们认为只有适当变量的个数才是合理的，如果变量太少，就会遗漏很多信息，导致模型不足以解释因变量，如果变量太多，就会导致信息重叠，同样导

致建模失真。目前国际上公认的比较合理的信息准则有很多种，所以研究者在选取滞后阶数时要适当加入自己的判断。在确定滞后阶数后，我们要确定协整秩，协整秩代表着协整关系的个数。变量之间往往会存在多个长期均衡关系，所以协整秩并不必然等于 1。在确定协整秩后，我们就可以构建相应的模型，并写出协整方程了。本例中，迹检验的结果如图 15.31 和图 15.32 所示。

```
. varsoc m d.s g

  Selection-order criteria
  Sample:  1983 - 2005                          Number of obs      =        23

  +---------------------------------------------------------------------------+
  |lag |    LL      LR      df    p      FPE       AIC      HQIC      SBIC    |
  |----+----------------------------------------------------------------------|
  |  0 | -298.833                      5.0e+07   26.2463   26.2836   26.3944  |
  |  1 | -263.187  71.291    9  0.000  5.0e+06*  23.9293   24.0783   24.5218* |
  |  2 | -255.196  15.982    9  0.067  5.7e+06   24.0171   24.2778   25.0538  |
  |  3 |   -245.8  18.793    9  0.027  6.3e+06   23.9826   24.3551   25.4637  |
  |  4 | -231.844  27.912*   9  0.001  5.3e+06   23.5516*  24.0359*   25.477  |
  +---------------------------------------------------------------------------+
   Endogenous:  m D.s g
    Exogenous:  _cons
```

图 15.31 根据信息准则确定变量滞后阶数

图 15.31 给出了根据信息准则确定的变量滞后阶数分析结果。最左列的 lag 表示的是滞后阶数，LL、LR 两列表示的是统计量，df 表示的是自由度，P 值表示的是对应滞后阶数下模型的显著性，FPE、AIC、HQIC、SBIC 代表的是 4 种信息准则，其中值越小越好，越应该选用，这一点也可以通过观察“*”号来验证，带“*”号的说明在本信息准则下的最优滞后阶数。最下面两行文字说明的是模型中的外生变量和内生变量，本例中，外生变量包括 m、D.s 、g（Endogenous: m D.s g），内生变量包括常数项（Exogenous:_cons）。

综上所述，可以看出选取滞后阶数为 1 阶或者 4 阶是比较合适的，但是为了使模型中的变量更多一些，更有说服力，我们选择滞后阶数为 4。

图 15.32 展示的是根据前面确定的滞后阶数确定协整秩的结果。分析本结果最直接的方式就是找到带有“*”号的迹统计量（Trace Statistic），本例中该值为 14.5747，对应的协整秩为 1，这说明本例中城乡人口净转移、城镇失业规模、城乡收入差距 3 个变量存在一个协整关系。

```
. vecrank m d.s g,lags(4)

                       Johansen tests for cointegration
Trend: constant                                         Number of obs =      23
Sample:  1983 - 2005                                             Lags =       4
-------------------------------------------------------------------------------
                                                             5%
maximum                                          trace    critical
  rank    parms       LL        eigenvalue    statistic     value
    0      30    -252.19968         .          40.7116      29.68
    1      35    -239.13121      0.67902       14.5747*     15.41
    2      38    -231.98625      0.46275        0.2848       3.76
    3      39    -231.84387      0.01230
-------------------------------------------------------------------------------
```

图 15.32 根据滞后阶数确定协整秩

至此，协整检验完毕。我们发现两种检验方法得到的结论是一致的。对于迹检验而言，同样可以构建出相应的模型来描述这种长期协整关系。这一点也放到本节的案例延伸部分来进行详细说明。

15.3.5 案例延伸

按照前面讲述的解决方法，可以对变量进行相应阶数的差分，然后进行回归，即可避免出现伪回归的情况。

1．EG-ADF 检验方法构建出的协整模型

如果假定 m 为因变量（真实情况需要进行格兰杰因果关系检验，将在下节中说明），则构建如下所示的模型方程：

$$d.m=a*d.g+b*d2.s+c*ecm_{t-1}+u$$

其中，a、b、c 为系数，ecm 为误差修正项，u 为误差扰动项。

ecm 误差修正项的模型方程为：

$$m=a*g+b*d.s+ecm_t$$

其中，a、b 为系数。实质上，ecm 是该模型方程的误差扰动项，或者说以 m 为因变量，以 g、d.s 为自变量进行最小二乘估计回归后的残差。

在上面的 EG-ADF 检验部分，得到的 ecm 模型方程为：

$$m=-1.229304*d1.s+793.4284*g-14.01591$$

该方程反映的是变量的长期均衡关系。

然后在主界面的“Command”文本框中首先输入命令：

```
regress d.m d2.s d.g l.e
```

并按键盘上的回车键进行确认，即可出现如图 15.33 所示的回归分析结果。

```
. regress d.m d2.s d.g l.e

      Source |       SS       df       MS              Number of obs =      26
-------------+------------------------------           F(  3,    22) =    1.67
       Model |  695996.067     3  231998.689           Prob > F      =  0.2021
    Residual |  3053062.28    22  138775.558           R-squared     =  0.1856
-------------+------------------------------           Adj R-squared =  0.0746
       Total |  3749058.34    25  149962.334           Root MSE      =  372.53

------------------------------------------------------------------------------
         D.m |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           s |
         D2. |   1.297896   2.025272     0.64   0.528     -2.90226    5.498052
             |
           g |
         D1. |   -26.2911   471.0633    -0.06   0.956    -1003.217    950.6345
             |
           e |
         L1. |  -.3580287   .1659561    -2.16   0.042    -.7022007   -.0138567
             |
       _cons |   27.56783   74.25575     0.37   0.714    -126.4292    181.5648
------------------------------------------------------------------------------
```

图 15.33　用 EG-ADF 检验方法构建协整模型

从上述分析结果中可以看到共有 26 个样本参与了分析。模型的 F 值(3, 22) ＝ 1.67，P 值（Prob > F）＝0.2021，说明模型整体上是差强人意的。模型的可决系数（R-squared）为 0.1856，

模型修正的可决系数（Adj R-squared）为 0.0746，说明模型解释能力偏弱。

模型的回归方程是：

d.m= 1.297896* d2.s– -26.2911*d1.g–0.3580287*l1.e+27.56783

变量 d2.s 的系数标准误是 2.025272，t 值为 0.64，P 值为 0.528，系数是非常不显著的，95%的置信区间为[-2.90226, 5.498052]。变量 d1.g 的系数标准误是 471.0633，t 值为-0.06，P 值为 0.956，系数也是非常不显著的，95%的置信区间为[-1003.217, 950.6345]。变量 l1.e 的系数标准误是 0.1659561，t 值为-2.16，P 值为 0.042，系数是比较显著的，95%的置信区间为[-0.7022007 , -0.0138567]。常数项的系数标准误是 74.25575，t 值为 0.37，P 值为 0.714，系数也是非常不显著的，95%的置信区间为[-126.4292 , 181.5648]。

2．迹检验方法构建出的协整模型

从上面的分析中可以看出，变量间的短期关系是非常不显著的，几乎没有什么关系。但是变量的长期均衡关系却很显著。下面利用另外一种更加精确的迹检验方法构建出的协整模型来详细研究变量间的这种长期均衡关系。

在进行迹检验完毕以后，在主界面的“Command”文本框中输入如下命令并按键盘上的回车键进行确认。

```
vec m d.s g,lags(4) rank(1)
```

即可得到如图 15.34~图 15.38 所示的分析结果。

```
. vec m d.s g,lags(4) rank(1)

Vector error-correction model

Sample:  1983 - 2005                          No. of obs      =        23
                                              AIC             =   23.8375
Log likelihood = -239.1312                    HQIC            =  24.27206
Det(Sigma_ml)  =    215429                    SBIC            =  25.56542

Equation           Parms      RMSE     R-sq      chi2     P>chi2
----------------------------------------------------------------
D_m                  11     317.064   0.6252   20.01941   0.0451
D2_s                 11     26.0643   0.7158   30.22438   0.0015
D_g                  11     .169976   0.4442   9.590791   0.5675
----------------------------------------------------------------
```

图 15.34 模型方程综述

图 15.34 说明的是分别把城乡人口净转移的一阶差分、城镇失业规模的二阶差分、城乡收入差距的一阶差分作为因变量时的模型方程综述。通过观察图 15.34 可以知道城乡人口净转移、城镇失业规模、城乡收入差距 3 个变量之间的协整关系可以通过 3 个方程来说明。此次值得强调的是，协整关系表示的仅仅是变量之间的某种长期联动关系，跟因果关系是毫无关联的，如果要探究变量之间的因果关系，换言之，就是确定让谁来作因变量的问题，就需要用到格兰杰因果关系检验，这种检验方法我们将在下一节中详细叙述。

本例中（实质上所有的协整关系都是一样的），3 个方程的样本情况（Sample: 1983 - 2005、No. of obs=23）、信息准则情况（AIC= 23.8375、HQIC= 24.27206、SBIC= 25.56542）等都是相同的。当把城乡人口净转移的一阶差分作为因变量时，模型的可决系数为 0.6252，卡方值是 20.01941，P 值为 0.0451；当把城镇失业规模的二阶差分作为因变量时，模型的可决系数为

0.7158，卡方值是 30.22428，P 值为 0.0015；当把城乡收入差距的一阶差分作为因变量时，模型的可决系数为 0.4442，卡方值是 9.590791，P 值为 0.5675。

图 15.35 展示的是把城乡人口净转移这一变量的一阶差分作为因变量时的方程模型具体情况。本分析结果的解析与一般的回归方程是一样的，前面多有介绍，限于篇幅不再赘述。

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
D_m						
_ce1						
L1.	.0055647	.0522526	0.11	0.915	-.0968486	.107978
m						
LD.	-.4071214	.2589529	-1.57	0.116	-.9146598	.1004169
L2D.	.1040884	.2985183	0.35	0.727	-.4809968	.6891736
L3D.	.3743418	.2320138	1.61	0.107	-.0803968	.8290804
s						
LD2.	-2.040869	2.395867	-0.85	0.394	-6.736682	2.654943
L2D2.	3.086168	2.368167	1.30	0.193	-1.555354	7.727691
L3D2.	-1.221802	2.495776	-0.49	0.624	-6.113433	3.66983
g						
LD.	-1030.141	553.9042	-1.86	0.063	-2115.774	55.49085
L2D.	-158.3343	679.8208	-0.23	0.816	-1490.758	1174.09
L3D.	1118.583	681.4178	1.64	0.101	-216.9715	2454.137
_cons	58.07797	99.26686	0.59	0.559	-136.4815	252.6374

图 15.35　以城乡人口净转移一阶差分为因变量

图 15.36 展示的是把城乡收入差距这一变量的一阶差分作为因变量时的方程模型具体情况。本分析结果的解析与一般的回归方程是一样的，前面多有介绍，限于篇幅不再赘述。

D2_s						
_ce1						
L1.	.0197186	.0042954	4.59	0.000	.0112997	.0281374
m						
LD.	.0306339	.0212872	1.44	0.150	-.0110883	.0723561
L2D.	.0523903	.0245397	2.13	0.033	.0042933	.1004872
L3D.	.0390845	.0190727	2.05	0.040	.0017027	.0764663
s						
LD2.	.3573081	.1969523	1.81	0.070	-.0287113	.7433275
L2D2.	.0424359	.1946753	0.22	0.827	-.3391206	.4239924
L3D2.	-.1436708	.2051654	-0.70	0.484	-.5457876	.2584459
g						
LD.	82.94072	45.53371	1.82	0.069	-6.303715	172.1852
L2D.	192.2813	55.88469	3.44	0.001	82.74937	301.8133
L3D.	155.86	56.01598	2.78	0.005	46.07073	265.6493
_cons	-16.38996	8.160235	-2.01	0.045	-32.38373	-.3961917

图 15.36　以城乡收入差距一阶差分为因变量

图 15.37 展示的是把城乡人口净转移这一变量的一阶差分作为因变量时的方程模型具体情况。本分析结果的解析与一般的回归方程是一样的，前面多有介绍，限于篇幅不再赘述。

D_g						
_ce1						
L1.	6.43e-06	.000028	0.23	0.818	-.0000485	.0000613
m						
LD.	-.000017	.0001388	-0.12	0.902	-.0002891	.0002551
L2D.	.0001119	.00016	0.70	0.484	-.0002017	.0004256
L3D.	.0000631	.0001244	0.51	0.612	-.0001807	.0003068
s						
LD2.	.0003646	.0012844	0.28	0.776	-.0021528	.002882
L2D2.	.0004478	.0012696	0.35	0.724	-.0020405	.0029361
L3D2.	-.0017889	.001338	-1.34	0.181	-.0044112	.0008335
g						
LD.	.1450003	.2969451	0.49	0.625	-.4370013	.727002
L2D.	.3762944	.3644483	1.03	0.302	-.3380111	1.0906
L3D.	-.037681	.3653045	-0.10	0.918	-.7536646	.6783026
_cons	.0299252	.0532164	0.56	0.574	-.0743771	.1342275

图 15.37 以城乡人口净转移一阶差分为因变量

图 15.38 展示的是本例 3 个变量间的协整方程。协整方程模型总体上是非常显著的，卡方值为 30.78642，P 值为 0.0000。

Cointegrating equations

Equation	Parms	chi2	P>chi2
_ce1	2	30.78462	0.0000

Identification: beta is exactly identified

Johansen normalization restriction imposed

beta	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
_ce1						
m	1	.	.	.	.	.
s						
D1.	-55.4957	13.60093	-4.08	0.000	-82.15303	-28.83837
g	-2005.838	1215.746	-1.65	0.099	-4388.657	376.981
_cons	2708.056	.	.	.	.	.

图 15.38 协整方程

协整方程的具体形式为：

m-55.4957d1.s-2005.838g+2708.056=0

如果把 m 作为因变量，对上面的等式进行变形，结果便是：

m=-2708.056+55.4957d1.s+2005.838g

可以发现 m 与 s、g 都是正向变动关系。这表示的含义是从长期来看，城乡人口净转移、城镇失业规模、城乡收入差距 3 个变量都是正向联动变动的。这个结论与对变量进行相应阶数差分后进行回归分析得到的结论不同，这个结论说明从长期来看，城镇失业规模和城乡人口净转移是正向变动的，这也是可以理解的，因为城乡人口净转移越多，城镇失业规模就有可能越大。而城镇失业规模越大，很可能也意味着城镇创造的就业机会越多，从而导致城乡人口净转移越大。

15.4 格兰杰因果关系检验

15.4.1 格兰杰因果关系检验的功能与意义

在 15.3 节中我们提到，协整关系表示的仅仅是变量之间的某种长期联动关系，跟因果关系是毫无关联的，如果要探究变量之间的因果关系，就需要用到格兰杰因果关系检验。格兰杰因果关系检验的基本思想是如果 A 变量是 B 变量的因，同时 B 变量不是 A 变量的因，那么 A 变量的滞后值就可以帮助预测 B 变量的未来值，同时 B 变量的滞后值却不能帮助预测 A 变量的未来值。这种思想反映到操作层面就是如果 A 变量是 B 变量的因，那么以 A 变量为因变量、以 A 变量的滞后值以及 B 变量的滞后值作为自变量进行最小二乘回归，则 B 变量的滞后值的系数显著。另外，需要强调 3 点：一是格兰杰因果关系并非真正意义的因果关系，表明的仅仅是数据上的一种动态相关关系，如果要准确界定变量的因果关系，需要相应的实践经验作为支撑；二是参与格兰杰因果关系检验的各变量要求是同阶单整的；三是存在协整关系的变量间至少有一种格兰杰因果关系。

15.4.2 相关数据来源

	下载资源:\video\chap15\…
	下载资源:\sample\chap15\案例15.dta

【例 15.4】本节沿用上节的案例，试通过格兰杰因果检验的方式来判断相关变量包括城乡人口净转移、城镇失业规模、城乡收入差距等变量之间的格兰杰因果关系。

15.4.3 Stata 分析过程

在前面几节中，我们通过单位根检验发现城乡人口净转移、城乡收入差距两个变量是一阶单整的，而城镇失业规模变量是二阶单整的，所以在进行格兰杰因果关系检验时选择的变量是：城乡人口净转移、城乡收入差距以及城镇失业规模的一阶差分。

格兰杰因果关系检验的操作步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中分别输入如下命令并按键盘上的回车键进行确认。

- regress m l.m dl.s：本命令旨在以 m 为因变量，以 l.m、dl.s 为自变量，进行最小二乘回归分析。
- test dl.s=0：本命令旨在检验变量 dl.s 系数的显著性。
- regress d.s dl.s l.m：本命令旨在以 d.s 为因变量，以 l.m、dl.s 为自变量，进行最小二乘回归分析。

- test l.m=0：本命令旨在检验变量 l.m 系数的显著性。
- regress m l.m l.g：本命令旨在以 m 为因变量，以 l.m、l.g 为自变量，进行最小二乘回归分析。
- test l.g=0：本命令旨在检验变量 l.g 系数的显著性。
- regress g l.g l.m：本命令旨在以 g 为因变量，以 l.m、l.g 为自变量，进行最小二乘回归分析。
- test l.m=0：本命令旨在检验变量 l.m 系数的显著性。
- regress g l.g dl.s：本命令旨在以 g 为因变量，以 l.g、dl.s 为自变量，进行最小二乘回归分析。
- test dl.s=0：本命令旨在检验变量 dl.s 系数的显著性。
- regress d.s dl.s l.g：本命令旨在以 d.s 为因变量，以 l.g、dl.s 为自变量，进行最小二乘回归分析。
- test l.g=0：本命令旨在检验变量 l.g 系数的显著性。

03 设置完毕后，等待输出结果。

15.4.4　结果分析

在 Stata 14.0 主界面的结果窗口可以看到如图 15.39~图 15.44 所示的分析结果。

图 15.39 展示的是城镇失业规模是否是城乡人口净转移的格兰杰因的检验结果。通过观察分析结果，可以看出 dl.s 的系数值是非常不显著的。具体体现在其 t 值、F 值以及 P 值上，关于这一结果的详细解读方法前面章节中多有提及，限于篇幅此处不再赘述，所以，我们可以比较有把握地得出结论，城镇失业规模不是城乡人口净转移的格兰杰因。

```
. regress m l.m dl.s

      Source |       SS           df       MS      Number of obs   =        26
-------------+----------------------------------   F(2, 23)        =     15.75
       Model |  4629469.26         2  2314734.63   Prob > F        =    0.0000
    Residual |  3380523.97        23  146979.303   R-squared       =    0.5780
-------------+----------------------------------   Adj R-squared   =    0.5413
       Total |  8009993.23        25  320399.729   Root MSE        =    383.38

------------------------------------------------------------------------------
           m |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           m |
         L1. |    .781863   .1432483     5.46   0.000     .4855314    1.078195
             |
           s |
         LD. |  -.0846817   1.601568    -0.05   0.958    -3.397777    3.228413
             |
       _cons |    275.103   176.1746     1.56   0.132    -89.34196    639.5479
------------------------------------------------------------------------------

. test dl.s=0

 ( 1)  LD.s = 0

       F(  1,    23) =    0.00
            Prob > F =    0.9583
```

图 15.39　城镇失业规模不是城乡人口净转移的格兰杰因

图 15.40 展示的是城乡人口净转移是否是城镇失业规模的格兰杰因的检验结果。通过观察

分析结果，可以看出 l.m 的系数值是非常不显著的。具体体现在其 t 值、F 值以及 P 值上，关于这一结果的详细解读方法前面章节中多有提及，限于篇幅此处不再赘述，所以，我们可以比较有把握地得出结论，城乡人口净转移不是城镇失业规模的格兰杰因。

```
. regress d.s dl.s l.m

      Source |       SS       df       MS              Number of obs =      26
-------------+------------------------------           F(  2,    23) =   10.60
       Model |  28844.9958     2  14422.4979           Prob > F      =  0.0005
    Residual |  31308.4809    23   1361.2383           R-squared     =  0.4795
-------------+------------------------------           Adj R-squared =  0.4343
       Total |  60153.4767    25  2406.13907           Root MSE      =  36.895

------------------------------------------------------------------------------
         D.s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           s |
         LD. |   .6456263    .154129     4.19   0.000     .3267863    .9644663
             |
           m |
         L1. |   .0115627   .0137857     0.84   0.410    -.0169552    .0400806
             |
       _cons |  -10.07413   16.95439    -0.59   0.558    -45.14697    24.99871
------------------------------------------------------------------------------

. test l.m=0

 ( 1)  L.m = 0

       F(  1,    23) =    0.70
            Prob > F =    0.4102
```

图 15.40　城乡人口净转移不是城镇失业规模的格兰杰因

图 15.41 展示的是城乡收入差距是否是城乡人口净转移的格兰杰因的检验结果。通过观察分析结果，可以看出 l.g 的系数值是非常不显著的。具体体现在其 t 值、F 值以及 P 值上，关于这一结果的详细解读方法前面章节中多有提及，限于篇幅此处不再赘述，所以，我们可以比较有把握地得出结论，城乡收入差距不是城乡人口净转移的格兰杰因。

```
. regress m l.m l.g

      Source |       SS       df       MS              Number of obs =      26
-------------+------------------------------           F(  2,    23) =   17.70
       Model |  4855190.69     2  2427595.35           Prob > F      =  0.0000
    Residual |  3154802.54    23  137165.328           R-squared     =  0.6061
-------------+------------------------------           Adj R-squared =  0.5719
       Total |  8009993.23    25  320399.729           Root MSE      =  370.36

------------------------------------------------------------------------------
           m |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           m |
         L1. |   .6777926    .156107     4.34   0.000     .3548607    1.000725
             |
           g |
         L1. |   272.6828   212.3726     1.28   0.212    -166.6435     712.009
             |
       _cons |  -7.728937   278.3084    -0.03   0.978    -583.4537    567.9958
------------------------------------------------------------------------------

. test l.g=0

 ( 1)  L.g = 0

       F(  1,    23) =    1.65
            Prob > F =    0.2119
```

图 15.41　城乡收入差距不是城乡人口净转移的格兰杰因

图 15.42 展示的是城乡人口净转移是否是城乡收入差距的格兰杰因的检验结果。通过观察

分析结果，可以看出 l.m 的系数值是非常不显著的。具体体现在其 t 值、F 值以及 P 值上，关于这一结果的详细解读方法前面章节中多有提及，限于篇幅此处不再赘述，所以，可以比较有把握地得出结论，城乡人口净转移不是城镇失业规模的格兰杰因。

```
. regress g  l.g l.m

      Source |       SS       df       MS              Number of obs =      26
-------------+------------------------------           F(  2,    23) =   65.41
       Model |  3.95900219     2  1.97950109           Prob > F      =  0.0000
    Residual |  .696013202    23  .030261444           R-squared     =  0.8505
-------------+------------------------------           Adj R-squared =  0.8375
       Total |  4.65501539    25  .186200615           Root MSE      =  .17396

------------------------------------------------------------------------------
           g |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           g |
         L1. |   .9152055   .0997519     9.17   0.000      .708853    1.121558
             |
           m |
         L1. |   .0000876   .0000733     1.19   0.244    -.0000641    .0002393
             |
       _cons |   .0514088   .1307221     0.39   0.698    -.2190104     .321828
------------------------------------------------------------------------------

. test l.m=0

 ( 1)  L.m = 0

       F(  1,    23) =    1.43
            Prob > F =    0.2443
```

图 15.42　城乡人口净转移不是城镇失业规模的格兰杰因

图 15.43 展示的是城镇失业规模是否是城乡收入差距的格兰杰因的检验结果。通过观察分析结果，可以看出 dl.s 的系数值是非常不显著的。具体体现在其 t 值、F 值以及 P 值上，关于这一结果的详细解读方法前面章节中多有提及，限于篇幅此处不再赘述，所以，可以比较有把握地得出结论，城镇失业规模是城乡收入差距的格兰杰因。

```
. regress g  l.g dl.s

      Source |       SS       df       MS              Number of obs =      26
-------------+------------------------------           F(  2,    23) =   74.99
       Model |  4.03608946     2  2.01804473           Prob > F      =  0.0000
    Residual |  .618925925    23  .026909823           R-squared     =  0.8670
-------------+------------------------------           Adj R-squared =  0.8555
       Total |  4.65501539    25  .186200615           Root MSE      =  .16404

------------------------------------------------------------------------------
           g |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           g |
         L1. |   .8465603   .1014616     8.34   0.000     .6366711     1.05645
             |
           s |
         LD. |    .001763   .0008338     2.11   0.046     .0000381    .0034879
             |
       _cons |   .2315428   .1468955     1.58   0.129    -.0723336    .5354193
------------------------------------------------------------------------------

. test dl.s=0

 ( 1)  LD.s = 0

       F(  1,    23) =    4.47
            Prob > F =    0.0455
```

图 15.43　城镇失业规模是城乡收入差距的格兰杰因

图 15.44 展示的是城乡收入差距是否是城镇失业规模的格兰杰因的检验结果。通过观察分析结果，可以看出 l.g 的系数值是非常不显著的。具体体现在其 t 值、F 值以及 P 值上，关于

这一结果的详细解读方法前面章节中多有提及，限于篇幅此处不再赘述，所以，可以比较有把握地得出结论，城乡收入差距不是城镇失业规模的格兰杰因。

```
. regress d.s dl.s l.g

      Source |       SS           df       MS      Number of obs   =        26
-------------+----------------------------------   F(  2,     23)  =     10.04
       Model |  28037.0225         2  14018.5112   Prob > F        =    0.0007
    Residual |  32116.4543        23  1396.36758   R-squared       =    0.4661
-------------+----------------------------------   Adj R-squared   =    0.4197
       Total |  60153.4767        25  2406.13907   Root MSE        =    37.368

------------------------------------------------------------------------------
         D.s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           s |
         LD. |    .714422   .1899441     3.76   0.001     .3214927    1.107351
             |
           g |
         L1. |   -7.56637   23.11245    -0.33   0.746    -55.37812    40.24538
             |
       _cons |   13.37976   33.46208     0.40   0.693    -55.84183    82.60135
------------------------------------------------------------------------------

. test l.g

 ( 1)  L.g = 0

       F(  1,    23) =    0.11
            Prob > F =    0.7463
```

图 15.44　城乡收入差距不是城镇失业规模的格兰杰因

综上所述，只有城镇失业规模是城乡收入差距的格兰杰因，其他变量之间均不存在格兰杰因果关系。当然，正如前面讲到的，格兰杰因果关系并不是真正的变量因果关系，变量实质的因果关系依靠有关理论或者实践经验的判断。格兰杰因果关系反映的仅仅是一种预测的效果，起到一种辅助的作用，所以，本例的格兰杰因果检验虽然没有得到预想的结果，但并不意味着模型的失败。读者们可以尝试增加其他更加有效的变量继续深入研究。

15.4.5　案例延伸

在前面的格兰杰因果关系检验的过程中，读者们可能会注意到我们使用的被假设为格兰杰因的自变量的滞后期均为 1 期。事实上可以多试几期，具体多少期读者可以根据研究的实际需要来加入自己的判断。例如，在检验城乡收入差距是否是城镇失业规模的格兰杰因的时候，可以把滞后期扩展为 5 期。在主界面的“Command”文本框中分别输入如下命令。

1．regress d.s dl.s l.g l2.g l3.g l4.g l5.g

本命令旨在以 d.s 为因变量，以 dl.s、l.g、l2.g、l3.g、l4.g、l5.g 为自变量，进行最小二乘回归分析。

2．test l.g=0

本命令旨在检验变量 l.g 系数的显著性。

3．test l2.g=0

本命令旨在检验变量 l2.g 系数的显著性。

4．test l3.g=0

本命令旨在检验变量 l3.g 系数的显著性。

5．test l4.g=0

本命令旨在检验变量 l4.g 系数的显著性。

6．test l5.g=0

本命令旨在检验变量 l5.g 系数的显著性。

按键盘上的回车键进行确认，即可出现如图 15.45 所示的分析检验结果。

```
. regress d.s dl.s l.g l2.g l3.g l4.g l5.g

      Source |       SS           df       MS      Number of obs   =        23
-------------+----------------------------------   F(6, 16)        =      2.93
       Model |  17451.8741         6  2908.64569   Prob > F        =    0.0402
    Residual |  15909.2876        16  994.330472   R-squared       =    0.5231
-------------+----------------------------------   Adj R-squared   =    0.3443
       Total |  33361.1617        22  1516.41644   Root MSE        =    31.533

------------------------------------------------------------------------------
         D.s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           s |
         LD. |   .3735399   .2580485     1.45   0.167    -.1734985    .9205782
             |
           g |
         L1. |   29.72947   52.48465     0.57   0.579    -81.53302     140.992
         L2. |   23.24441   63.76133     0.36   0.720    -111.9236    158.4124
         L3. |  -21.91515   58.52375    -0.37   0.713      -145.98    102.1497
         L4. |  -62.81455   62.25527    -1.01   0.328    -194.7898    69.16072
         L5. |   26.73216   49.88799     0.54   0.599    -79.02566      132.49
             |
       _cons |   18.56089   32.80009     0.57   0.579     -50.9722    88.09398
------------------------------------------------------------------------------

. test l.g=l2.g=l3.g=l4.g=l5.g=0

 ( 1)  L.g - L2.g = 0
 ( 2)  L.g - L3.g = 0
 ( 3)  L.g - L4.g = 0
 ( 4)  L.g - L5.g = 0
 ( 5)  L.g = 0

       F(  5,    16) =    0.69
            Prob > F =    0.6376
```

图 15.45　分析结果图

通过观察分析结果，可以看出 l.g、l2.g、l3.g、l4.g、l5.g 的系数值都是非常不显著的。具体体现在其 t 值、F 值以及 P 值上，关于这一结果的详细解读方法前面章节中多有提及，限于篇幅此处不再赘述，所以，我们可以比较有把握地得出结论，城乡收入差距不是城镇失业规模的格兰杰因。其他变量间的检验是类似的，读者可以自己尝试分析。

15.5　本章习题

某公司自 1983 年成立以来，主要的经营指标数据包括年销售收入、年运营成本、母公司考核系数等，如表 15.3 所示。试将数据整理成 Stata 数据文件，并进行以下操作。

（1）定义时间序列，并绘制各时间序列变量的时间趋势图，进行简要分析。

（2）试通过单位根检验的方式来判断相关变量，包括年销售收入、年运营成本、母公司

考核系数等变量是否平稳。

（3）试通过 EG-ADF 检验、迹检验等两种协整检验的方式来判断相关变量，包括年销售收入、年运营成本、母公司考核系数等变量是否存在长期协整关系。

（4）试通过格兰杰因果检验的方式来判断相关变量，包括年销售收入、年运营成本、母公司考核系数变量之间的格兰杰因果关系。

表 15.3　某公司经营指标数据及相关变量数据

年份	年销售收入/万元	年运营成本/万元	母公司考核系数
1983	943.77	264.4	1.5
1984	1101.69	276.6	1.24
1985	484.28	296.2	0.98
1986	814.63	439.5	1.24
1987	1055.05	349.4	0.98
1988	571.68	271.4	0.82
…	…	…	…
2008	1821.55	800	2.23
2009	1779.12	827	2.21
2010	1785.18	839	2.22
2011	1834.26	476.4	1.47
2012	1832.07	519.6	1.51

第 16 章　Stata 面板数据分析

面板数据（Panel Data）又被称为平行数据，指的是对某变量在一定时间段内持续跟踪观测的结果。面板数据兼具了横截面数据和时间序列数据的特点，既有横截面维度（在同一时间段内有多个观测样本），又有时间序列维度（同一样本在多个时间段内被观测到）。面板数据通常样本数量相对较多，也可以有效解决遗漏变量的问题，还可以提供更多样本动态行为的信息，具有横截面数据和时间序列数据无可比拟的优势。根据横截面维度和时间序列维度相对长度的大小，面板数据被区分为长面板数据和短面板数据。下面就来一一介绍这两种面板数据分析方法在实例中的应用。

16.1　实例一——短面板数据分析

16.1.1　短面板数据分析的功能与意义

短面板数据是面板数据的一种，其主要特征是横截面维度比较大而时间维度相对较小，或者说，同一期间内被观测的个体数量较多而被观测的期间较少。短面板数据分析方法包括直接最小二乘回归分析、固定效应回归分析、随机效应回归分析、组间估计量回归分析等多种。下面就以实例的方式来介绍一下这几种方法的具体应用。

16.1.2　相关数据来源

	下载资源:\video\chap16\…
	下载资源:\sample\chap16\案例16.1.dta

【例 16.1】 A 公司是一家销售饮料的连锁公司，经营范围遍布全国 20 个省市，各省市连锁店 2008－2012 年的相关销售数据（包括销售收入、促销费用以及创造利润等数据）如表 16.1 所示。试用多种短面板数据回归分析方法深入研究销售量和促销费用对创造利润的影响关系。

表 16.1　A 公司各省市连锁店销售收入、促销费用以及创造利润数据（2008－2012 年）

年份	销售收入/万元	促销费用/万元	创造利润/万元	地区
2008	256	13.28039	12.47652	北京
2009	289	12.88284	12.1826	北京
2010	321	12.86566	12.26754	北京
2011	135	13.166	12.25672	北京
2012	89	13.01277	12.21607	北京

（续表）

年份	销售收入/万元	促销费用/万元	创造利润/万元	地区
2008	159	11.00874	9.236008	天津
…	…	…	…	…
2012	226.0475	10.77687	10.39666	甘肃
2008	229.2657	11.41421	10.47813	青海
2009	228.9225	11.10796	10.19802	青海
2010	229.2313	11.36674	10.47249	青海
2011	229.0406	11.1375	10.22485	青海
2012	229.1517	11.24112	10.30762	青海

16.1.3 Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有 5 个变量，分别是年份、销售收入、促销费用、创造利润以及地区。我们把年份变量定义为 year，把销售收入变量定义为 sale，把促销费用变量定义为 cost，把创造利润变量定义为 profit，把地区变量定义为 diqu。变量类型及长度采取系统默认方式，然后录入相关数据。相关操作在第 1 章中已有详细讲述。录入完成后数据如图 16.1 所示。

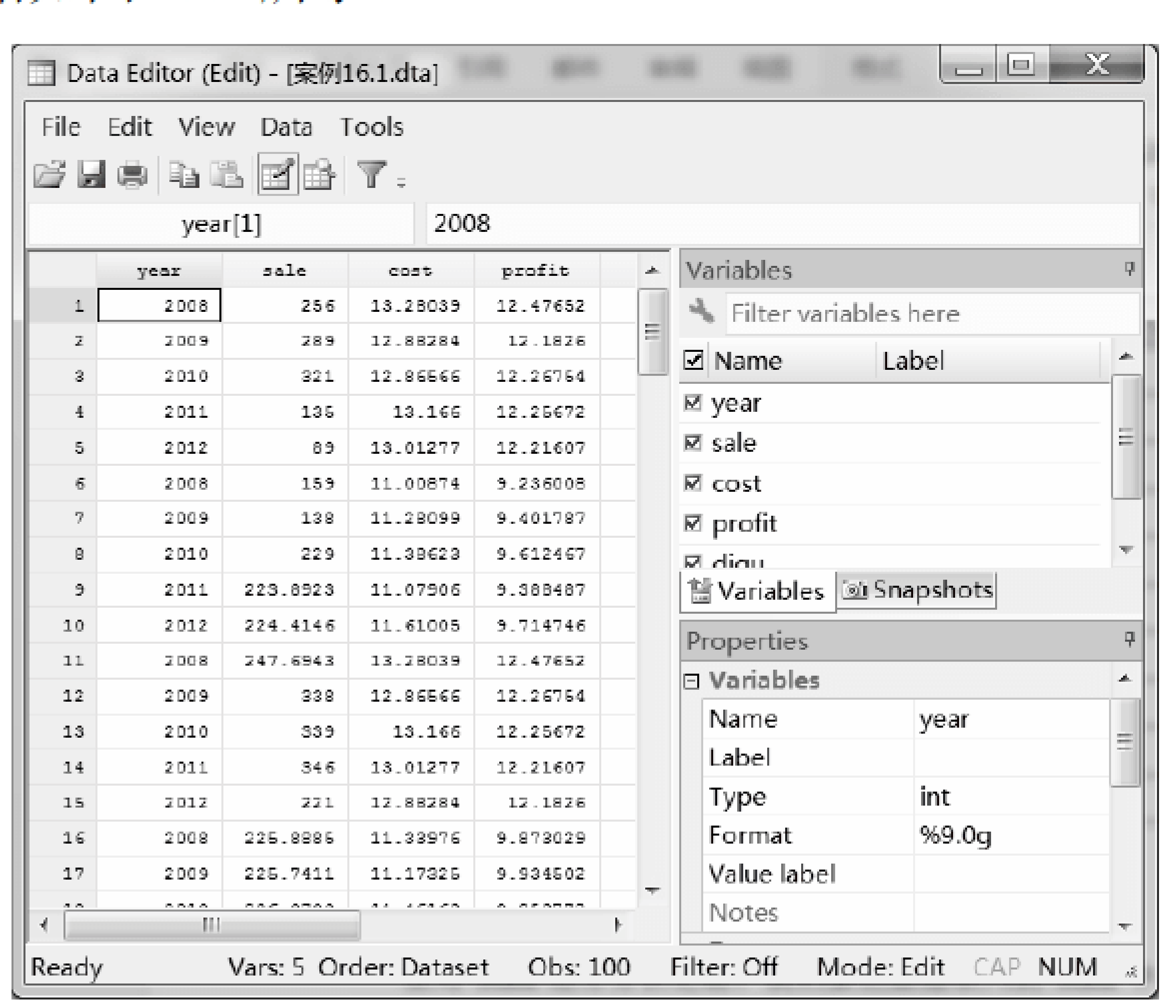


图 16.1　案例 16.1 数据

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入如下命令。

- list year sale cost profit：本命令的含义是对 4 个变量所包含的样本数据进行一一展示，以便简单直观地观测出数据的具体特征，为深入分析做好必要准备。
- encode diqu,gen(region)：因为面板数据要求其中的个体变量取值必须为整数而且不允许有重复，所以需要对各个观测样本进行有序编号。本命令旨在将 diqu 这一字符串变量转化为数值型变量，以便进行下一步操作。

- xtset region year：本命令的含义是对面板数据进行定义，其中横截面维度变量为上步生成的 region，时间序列变量为 year。
- xtdes：本命令旨在观测面板数据的结构，考察面板数据特征，为后续分析做好必要准备。
- xtsum：本命令旨在显示面板数据组内、组间以及整体的统计指标。
- xttab sale：本命令旨在显示“sale”变量组内、组间以及整体的分布频率。
- xttab cost：本命令旨在显示“cost”变量组内、组间以及整体的分布频率。
- xttab profit：本命令旨在显示“profit”变量组内、组间以及整体的分布频率。
- xtline sale：本命令旨在对每个个体显示“sale”变量的时间序列图。
- xtline cost：本命令旨在对每个个体显示“cost”变量的时间序列图。
- xtline profit：本命令旨在对每个个体显示“profit”变量的时间序列图。
- reg profit sale cost：本命令的含义是以 profit 为因变量，以 sale、cost 为自变量，进行最小二乘回归分析。
- reg profit sale cost,vce(cluster region)：本命令的含义是以 profit 为因变量，以 sale、cost 为自变量，并使用以“region”为聚类变量的聚类稳健标准差，进行最小二乘回归分析。
- xtreg profit sale cost,fe vce(cluster region)：本命令的含义是以 profit 为因变量，以 sale、cost 为自变量，并使用以“region”为聚类变量的聚类稳健标准差，进行固定效应回归分析。
- xtreg profit sale cost,fe：本命令的含义是以 profit 为因变量，以 sale、cost 为自变量，进行固定效应回归分析。
- estimates store fe：本命令的含义是存储固定效应回归分析的估计结果。
- xi:xtreg　profit sale cost　i.region,vce(cluster region)：本命令旨在通过构建最小二乘虚拟变量模型来分析固定效应模型是否优于最小二乘回归分析。
- tab year,gen(year)：本命令旨在创建年度变量的多个虚拟变量。
- xtreg profit sale cost year2-year5,fe vce(cluster region)：本命令旨在通过构建双向固定效应模型来检验模型中是否应该包含时间效应。
- test year2 year3 year4 year5：本命令的含义是在上步回归的基础上，通过测试各虚拟变量的系数联合显著性来检验是否应该在模型中纳入时间效应。
- xtreg profit sale cost,re vce(cluster region)：本命令的含义是以 profit 为因变量，以 sale、cost 为自变量，并使用以“region”为聚类变量的聚类稳健标准差，进行随机效应回归分析。
- xttest0：本命令的含义是在上步回归的基础上，进行假设检验来判断随机效应模型是否优于最小二乘回归模型。
- xtreg profit sale cost,mle：本命令的含义是以 profit 为因变量，以 sale、cost 为自变量，并使用最大似然估计方法，进行随机效应回归分析。
- xtreg profit sale cost,be：本命令的含义是以 profit 为因变量，以 sale、cost 为自变量，并使用组间估计量，进行组间估计量回归分析。

03 设置完毕后，按键盘上的回车键，等待输出结果。

16.1.4 结果分析

在 Stata 14.0 主界面的结果窗口可以看到如图 16.2~图 16.25 所示的分析结果。

图 16.2 是对数据进行展示的结果。它的目的是通过对变量所包含的样本数据进行一一展示，以便简单直观地观测出数据的具体特征，为深入分析做好必要准备。

```
. list year sale cost profit
```

	year	sale	cost	profit
1.	2008	256	13.28039	12.47652
2.	2009	289	12.88284	12.1826
3.	2010	321	12.86566	12.26754
4.	2011	135	13.166	12.25672
5.	2012	89	13.01277	12.21607
6.	2008	159	11.00874	9.236008
7.	2009	138	11.28099	9.401787
8.	2010	229	11.38623	9.612467
9.	2011	223.8923	11.07906	9.388487
10.	2012	224.4146	11.61005	9.714746
11.	2008	247.6943	13.28039	12.47652
12.	2009	338	12.86566	12.26754
13.	2010	339	13.166	12.25672
14.	2011	346	13.01277	12.21607
15.	2012	221	12.88284	12.1826
16.	2008	225.8885	11.33976	9.873029
17.	2009	225.7411	11.17325	9.934502
18.	2010	226.0703	11.46163	9.853772
19.	2011	225.9849	11.42737	9.879707
20.	2012	225.4703	11.12873	9.864227
21.	2008	223.664	10.86284	10.06305
22.	2009	223.3596	10.7579	9.720165
23.	2010	189	11.32298	9.786392
24.	2011	194	11.32055	9.804219
25.	2012	191	11.19272	9.89948
26.	2008	229.834	11.60368	10.15619
27.	2009	229.5091	11.48143	10.18036
28.	2010	229.6875	11.51192	10.05277
29.	2011	229.9539	11.86005	10.35711
30.	2012	229.9492	11.73527	10.28637
31.	2008	195	11.32298	9.786392
32.	2009	190	10.7579	9.720165
33.	2010	196	11.19272	9.89948
34.	2011	191	11.32055	9.804219
35.	2012	223.664	10.86284	10.06305
36.	2008	230.2526	11.35158	10.38807
37.	2009	230.4395	11.65529	10.57132
38.	2010	230.1745	11.30836	10.52889
39.	2011	230.3779	11.48555	10.59037
40.	2012	230.4235	11.59451	10.56721
41.	2008	224.4761	10.83762	10.16969
42.	2009	224.5877	10.9682	10.13896
43.	2010	224.7289	11.18164	10.32286
44.	2011	224.373	10.77896	10.34432
45.	2012	224.7235	11.10796	10.17884
46.	2008	228.9225	11.10796	10.19802
47.	2009	229.2313	11.36674	10.47249
48.	2010	229.2657	11.41421	10.47813
49.	2011	229.1517	11.24112	10.30762
50.	2012	229.0406	11.1375	10.22485
51.	2008	224.4039	11.38623	9.612467
52.	2009	224.2034	11.28099	9.401787
53.	2010	223.8923	11.07906	9.388487
54.	2011	224.4146	11.61005	9.714746
55.	2012	223.5251	11.00874	9.236008
56.	2008	226.2307	10.91509	10.51732
57.	2009	226.1334	10.80771	10.43588
58.	2010	226.4084	11.14041	10.55451
59.	2011	226.3114	11.0021	10.4631
60.	2012	226.0475	10.77687	10.39666
61.	2008	230.4395	11.65529	10.57132
62.	2009	230.2526	11.35158	10.38807
63.	2010	230.1745	11.30836	10.52889
64.	2011	230.4235	11.59451	10.56721
65.	2012	230.3779	11.48555	10.59037
66.	2008	224.373	10.77896	10.34432
67.	2009	224.7235	11.10796	10.17884
68.	2010	224.7289	11.18164	10.32286
69.	2011	224.5877	10.9682	10.13896
70.	2012	224.4761	10.83762	10.16969
71.	2008	231.01	11.6994	9.914922
72.	2009	231.6112	11.89614	10.15891
73.	2010	231.4499	11.82188	10.15774
74.	2011	231.233	11.73847	10.01055
75.	2012	231.7159	12.09234	10.28739
76.	2008	229.6875	11.51192	10.05277
77.	2009	229.5091	11.48143	10.18036
78.	2010	229.9539	11.86005	10.35711
79.	2011	229.9492	11.73527	10.28637
80.	2012	229.834	11.60368	10.15619
81.	2008	201	11.17325	9.934502
82.	2009	198	11.33976	9.873029
83.	2010	199	11.46163	9.853772
84.	2011	201	11.12873	9.864227
85.	2012	201	11.42737	9.879707
86.	2008	198	11.6994	9.914922
87.	2009	231.6112	11.89614	10.15891
88.	2010	231.7159	12.09234	10.28739
89.	2011	231.4499	11.82188	10.15774
90.	2012	231.233	11.73847	10.01055
91.	2008	226.4084	11.14041	10.55451
92.	2009	226.3114	11.0021	10.4631
93.	2010	226.2307	10.91509	10.51732
94.	2011	226.1334	10.80771	10.43588
95.	2012	226.0475	10.77687	10.39666
96.	2008	229.2657	11.41421	10.47813
97.	2009	228.9225	11.10796	10.19802
98.	2010	229.2313	11.36674	10.47249
99.	2011	229.0406	11.1375	10.22485
100.	2012	229.1517	11.24112	10.30762

图 16.2　展示数据

从如图 16.2 所示的分析结果中可以看出，数据的总体质量还是可以的，没有极端异常值，变量间的量纲差距也是可以接受的，可以进入下一步的分析。

图 16.3 是将 diqu 这一字符串变量转化为数值型变量 region 的结果。选择“Data” | “Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 16.3 所示的变量 region 的相关数据。

图 16.4 为对面板数据进行定义的结果，其中横截面维度变量为上步生成的 region，时间序列变量为 year。

	year	sale	cost	profit	diqu	region
1	2008	256	13.28039	12.47652	北京	北京
2	2009	289	12.88284	12.1826	北京	北京
3	2010	321	12.86566	12.26754	北京	北京
4	2011	135	13.166	12.25672	北京	北京
5	2012	89	13.01277	12.21607	北京	北京
6	2008	226.4084	11.14041	10.55451	甘肃	甘肃
7	2009	226.3114	11.0021	10.4631	甘肃	甘肃
8	2010	226.2307	10.91509	10.51732	甘肃	甘肃
9	2011	226.1334	10.80771	10.43588	甘肃	甘肃
10	2012	226.0475	10.77687	10.39666	甘肃	甘肃
11	2008	224.4039	11.38623	9.612467	广东	广东
12	2009	224.2034	11.28099	9.401787	广东	广东
13	2010	223.8923	11.07906	9.388487	广东	广东
14	2011	224.4146	11.61005	9.714746	广东	广东
15	2012	223.5251	11.00874	9.236008	广东	广东
16	2008	226.2307	10.91509	10.51732	广西	广西
17	2009	226.1334	10.80771	10.43588	广西	广西
18	2010	226.4084	11.14041	10.55451	广西	广西
19	2011	226.3114	11.0021	10.4631	广西	广西
20	2012	226.0475	10.77687	10.39666	广西	广西
21	2008	230.4395	11.65529	10.57132	贵州	贵州
22	2009	230.2526	11.35158	10.38807	贵州	贵州
23	2010	230.1745	11.30836	10.52889	贵州	贵州
24	2011	230.4235	11.59451	10.56721	贵州	贵州
25	2012	230.3779	11.48555	10.59037	贵州	贵州
26	2008	201	11.17325	9.934502	海南	海南
27	2009	198	11.33976	9.873029	海南	海南
28	2010	199	11.46163	9.853772	海南	海南

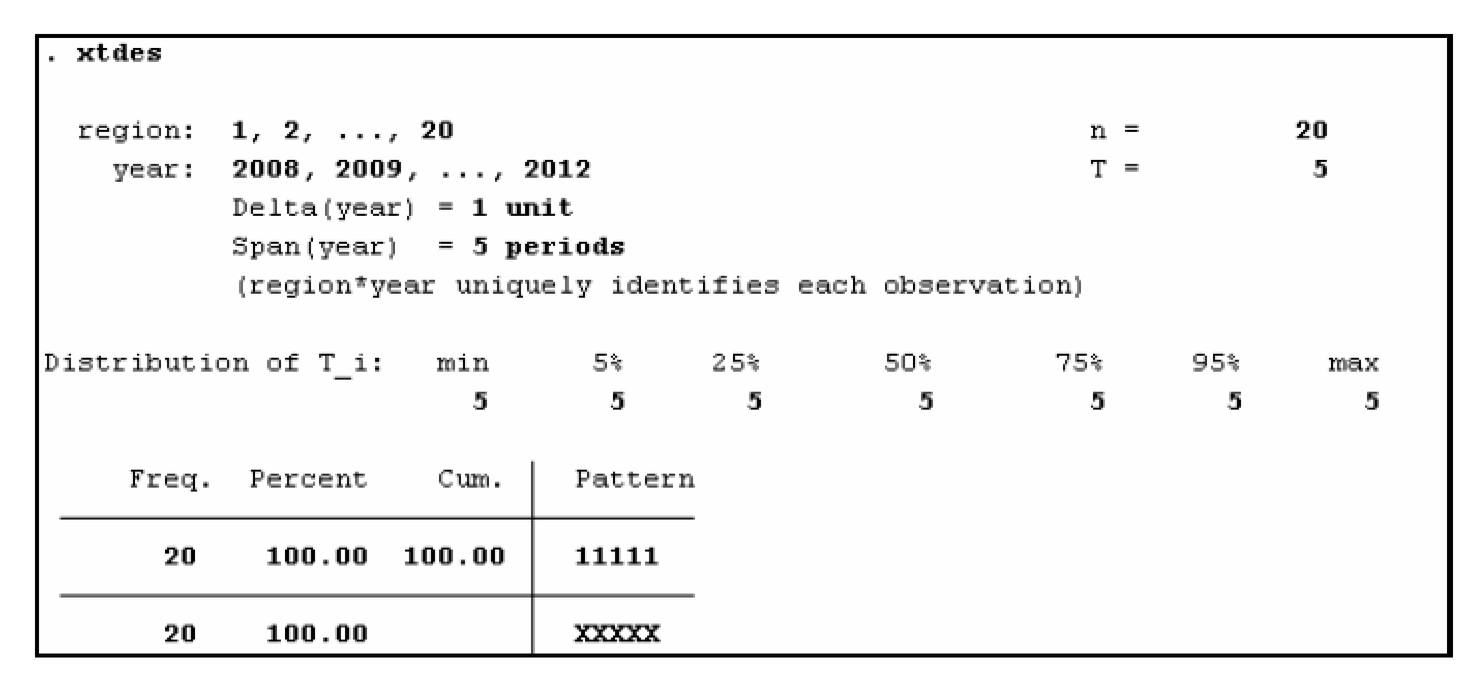
图 16.3　region 的相关数据

```
xtset region year
       panel variable:  region (strongly balanced)
        time variable:  year, 2008 to 2012
                delta:  1 unit
```

图 16.4　对面板数据进行定义

从图 16.4 中可以看出这是一个平衡的面板数据。

图 16.5 是面板数据结构的结果。



```
. xtdes

  region:  1, 2, ..., 20                                   n =         20
    year:  2008, 2009, ..., 2012                           T =          5
           Delta(year) = 1 unit
           Span(year)  = 5 periods
           (region*year uniquely identifies each observation)

Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                         5       5       5         5         5       5       5

     Freq.  Percent    Cum. |  Pattern
 ---------------------------+---------
       20    100.00  100.00 |  11111
 ---------------------------+---------
       20    100.00         |  XXXXX
```

图 16.5　面板数据结构

从图 16.5 可以看出该面板数据的横截面维度 region 为 1~20 共 20 个取值，时间序列维度 year 为 2008~2012 共 5 个取值，属于短面板数据，而且观测样本在时间上的分布也非常均匀。

图 16.6 是面板数据组内、组间以及整体的统计指标的结果。

在短面板数据中，同一时间段内的不同观测样本构成一个组。从图 16.6 中可以看出，变量 year 的组间标准差是 0，因为不同组的这一变量取值完全相同，同时变量 region 的组内标准差也为 0，因为分布在同一组的数据属于同一个地区。

```
. xtsum

Variable         |      Mean   Std. Dev.       Min        Max |    Observations
-----------------+--------------------------------------------+----------------
year     overall |      2010   1.421338       2008       2012 |     N =     100
         between |                    0       2010       2010 |     n =      20
         within  |             1.421338       2008       2012 |     T =       5
                 |                                            |
sale     overall |  225.0378   32.75807         89        346 |     N =     100
         between |             20.83152   194.8614   298.3389 |     n =      20
         within  |             25.62562   96.03781   328.0378 |     T =       5
                 |                                            |
cost     overall |  11.48361   .6108847    10.7579   13.28039 |     N =     100
         between |             .6012933   10.92844   13.04153 |     n =      20
         within  |             .1619716   11.15011   11.82065 |     T =       5
                 |                                            |
profit   overall |  10.33686   .7258455   9.236008   12.47652 |     N =     100
         between |             .7329161   9.470699   12.27989 |     n =      20
         within  |             .1067208   10.10217    10.5809 |     T =       5
                 |                                            |
diqu     overall |         .          .          .          . |     N =       0
         between |                    .          .          . |     n =       0
         within  |                    .          .          . |     T =       .
                 |                                            |
region   overall |      10.5   5.795331          1         20 |     N =     100
         between |              5.91608          1         20 |     n =      20
         within  |                    0       10.5       10.5 |     T =       5
```

图 16.6　面板数据统计指标

图 16.7 是“sale”变量组内、组间以及整体的分布频率的结果。

. xttab sale

sale	Overall Freq.	Overall Percent	Between Freq.	Between Percent	Within Percent
89	1	1.00	1	5.00	20.00
135	1	1.00	1	5.00	20.00
138	1	1.00	1	5.00	20.00
159	1	1.00	1	5.00	20.00
189	1	1.00	1	5.00	20.00
190	1	1.00	1	5.00	20.00
191	2	2.00	2	10.00	20.00
194	1	1.00	1	5.00	20.00
195	1	1.00	1	5.00	20.00
196	1	1.00	1	5.00	20.00
198	2	2.00	2	10.00	20.00
199	1	1.00	1	5.00	20.00
201	3	3.00	1	5.00	60.00
221	1	1.00	1	5.00	20.00
223.3596	1	1.00	1	5.00	20.00
223.5251	1	1.00	1	5.00	20.00
223.664	2	2.00	2	10.00	20.00
223.8923	2	2.00	2	10.00	20.00
224.2034	1	1.00	1	5.00	20.00
224.373	2	2.00	2	10.00	20.00
224.4039	1	1.00	1	5.00	20.00
224.4146	2	2.00	2	10.00	20.00
224.4761	2	2.00	2	10.00	20.00
224.5877	2	2.00	2	10.00	20.00
224.7235	2	2.00	2	10.00	20.00
224.7289	2	2.00	2	10.00	20.00
225.4703	1	1.00	1	5.00	20.00
225.7411	1	1.00	1	5.00	20.00
225.8885	1	1.00	1	5.00	20.00
225.9849	1	1.00	1	5.00	20.00
226.0475	2	2.00	2	10.00	20.00
226.0703	1	1.00	1	5.00	20.00
226.1334	2	2.00	2	10.00	20.00
226.2307	2	2.00	2	10.00	20.00
226.3114	2	2.00	2	10.00	20.00
226.4084	2	2.00	2	10.00	20.00
228.9225	2	2.00	2	10.00	20.00
229	1	1.00	1	5.00	20.00
229.0406	2	2.00	2	10.00	20.00
229.1517	2	2.00	2	10.00	20.00
229.2313	2	2.00	2	10.00	20.00
229.2657	2	2.00	2	10.00	20.00
229.5091	2	2.00	2	10.00	20.00
229.6875	2	2.00	2	10.00	20.00
229.834	2	2.00	2	10.00	20.00
229.9492	2	2.00	2	10.00	20.00
229.9539	2	2.00	2	10.00	20.00
230.1745	2	2.00	2	10.00	20.00
230.2526	2	2.00	2	10.00	20.00
230.3779	2	2.00	2	10.00	20.00
230.4235	2	2.00	2	10.00	20.00
230.4395	2	2.00	2	10.00	20.00
231.01	1	1.00	1	5.00	20.00
231.233	2	2.00	2	10.00	20.00
231.4499	2	2.00	2	10.00	20.00
231.6112	2	2.00	2	10.00	20.00
231.7159	2	2.00	2	10.00	20.00
247.6943	1	1.00	1	5.00	20.00
256	1	1.00	1	5.00	20.00
289	1	1.00	1	5.00	20.00
321	1	1.00	1	5.00	20.00
338	1	1.00	1	5.00	20.00
339	1	1.00	1	5.00	20.00
346	1	1.00	1	5.00	20.00
Total	100	100.00	98 (n = 20)	490.00	20.41

图 16.7　“sale”变量组内、组间以及整体的分布频率

图 16.8 是“cost”变量组内、组间以及整体的分布频率的结果。

. xttab cost

cost	Overall Freq.	Overall Percent	Between Freq.	Between Percent	Within Percent
10.7579	2	2.00	2	10.00	20.00
10.77687	2	2.00	2	10.00	20.00
10.77896	2	2.00	2	10.00	20.00
10.80771	2	2.00	2	10.00	20.00
10.83762	2	2.00	2	10.00	20.00
10.86284	2	2.00	2	10.00	20.00
10.91509	2	2.00	2	10.00	20.00
10.9682	2	2.00	2	10.00	20.00
11.0021	2	2.00	2	10.00	20.00
11.00874	2	2.00	2	10.00	20.00
11.07906	2	2.00	2	10.00	20.00
11.10796	4	4.00	4	20.00	20.00
11.12873	2	2.00	2	10.00	20.00
11.1375	2	2.00	2	10.00	20.00
11.14041	2	2.00	2	10.00	20.00
11.17325	2	2.00	2	10.00	20.00
11.18164	2	2.00	2	10.00	20.00
11.19272	2	2.00	2	10.00	20.00
11.24112	2	2.00	2	10.00	20.00
11.28099	2	2.00	2	10.00	20.00
11.30836	2	2.00	2	10.00	20.00
11.32055	2	2.00	2	10.00	20.00
11.32298	2	2.00	2	10.00	20.00
11.33976	2	2.00	2	10.00	20.00
11.35158	2	2.00	2	10.00	20.00
11.36674	2	2.00	2	10.00	20.00
11.38623	2	2.00	2	10.00	20.00
11.41421	2	2.00	2	10.00	20.00
11.42737	2	2.00	2	10.00	20.00
11.46163	2	2.00	2	10.00	20.00
11.48143	2	2.00	2	10.00	20.00
11.48555	2	2.00	2	10.00	20.00
11.51192	2	2.00	2	10.00	20.00
11.59451	2	2.00	2	10.00	20.00
11.60368	2	2.00	2	10.00	20.00
11.61005	2	2.00	2	10.00	20.00
11.65529	2	2.00	2	10.00	20.00
11.6994	2	2.00	2	10.00	20.00
11.73527	2	2.00	2	10.00	20.00
11.73847	2	2.00	2	10.00	20.00
11.82188	2	2.00	2	10.00	20.00
11.86005	2	2.00	2	10.00	20.00
11.89614	2	2.00	2	10.00	20.00
12.09234	2	2.00	2	10.00	20.00
12.86566	2	2.00	2	10.00	20.00
12.88284	2	2.00	2	10.00	20.00
13.01277	2	2.00	2	10.00	20.00
13.166	2	2.00	2	10.00	20.00
13.28039	2	2.00	2	10.00	20.00
Total	100	100.00	100 (n = 20)	500.00	20.00

图 16.8　“cost”变量组内、组间以及整体的分布频率

图 16.9 是“profit”变量组内、组间以及整体的分布频率的结果。

. xttab profit

profit	Overall Freq.	Overall Percent	Between Freq.	Between Percent	Within Percent
9.236008	2	2.00	2	10.00	20.00
9.388487	2	2.00	2	10.00	20.00
9.401787	2	2.00	2	10.00	20.00
9.612467	2	2.00	2	10.00	20.00
9.714746	2	2.00	2	10.00	20.00
9.720165	2	2.00	2	10.00	20.00
9.786392	2	2.00	2	10.00	20.00
9.804219	2	2.00	2	10.00	20.00
9.853772	2	2.00	2	10.00	20.00
9.864227	2	2.00	2	10.00	20.00
9.873029	2	2.00	2	10.00	20.00
9.879707	2	2.00	2	10.00	20.00
9.89948	2	2.00	2	10.00	20.00
9.914922	2	2.00	2	10.00	20.00
9.934502	2	2.00	2	10.00	20.00
10.01055	2	2.00	2	10.00	20.00
10.05277	2	2.00	2	10.00	20.00
10.06305	2	2.00	2	10.00	20.00
10.13896	2	2.00	2	10.00	20.00
10.15619	2	2.00	2	10.00	20.00
10.15774	2	2.00	2	10.00	20.00
10.15891	2	2.00	2	10.00	20.00
10.16969	2	2.00	2	10.00	20.00
10.17884	2	2.00	2	10.00	20.00
10.18036	2	2.00	2	10.00	20.00
10.19802	2	2.00	2	10.00	20.00
10.22485	2	2.00	2	10.00	20.00
10.28637	2	2.00	2	10.00	20.00
10.28739	2	2.00	2	10.00	20.00
10.30762	2	2.00	2	10.00	20.00
10.32286	2	2.00	2	10.00	20.00
10.34432	2	2.00	2	10.00	20.00
10.35711	2	2.00	2	10.00	20.00
10.38807	2	2.00	2	10.00	20.00
10.39665	2	2.00	2	10.00	20.00
10.43588	2	2.00	2	10.00	20.00
10.4631	2	2.00	2	10.00	20.00
10.47249	2	2.00	2	10.00	20.00
10.47813	2	2.00	2	10.00	20.00
10.51732	2	2.00	2	10.00	20.00
10.52889	2	2.00	2	10.00	20.00
10.55451	2	2.00	2	10.00	20.00
10.56721	2	2.00	2	10.00	20.00
10.57132	2	2.00	2	10.00	20.00
10.59037	2	2.00	2	10.00	20.00
12.1826	2	2.00	2	10.00	20.00
12.21607	2	2.00	2	10.00	20.00
12.25672	2	2.00	2	10.00	20.00
12.26754	2	2.00	2	10.00	20.00
12.47652	2	2.00	2	10.00	20.00
Total	100	100.00	100 (n = 20)	500.00	20.00

图 16.9　“profit”变量组内、组间以及整体的分布频率

图 16.10 是对每个个体显示“sale”变量的时间序列图的结果。

从图 16.10 可以看出，不同地区的销售收入的时间趋势是不一致的，有的地区变化非常平稳，有的地区先升后降，有的地区先降后升。

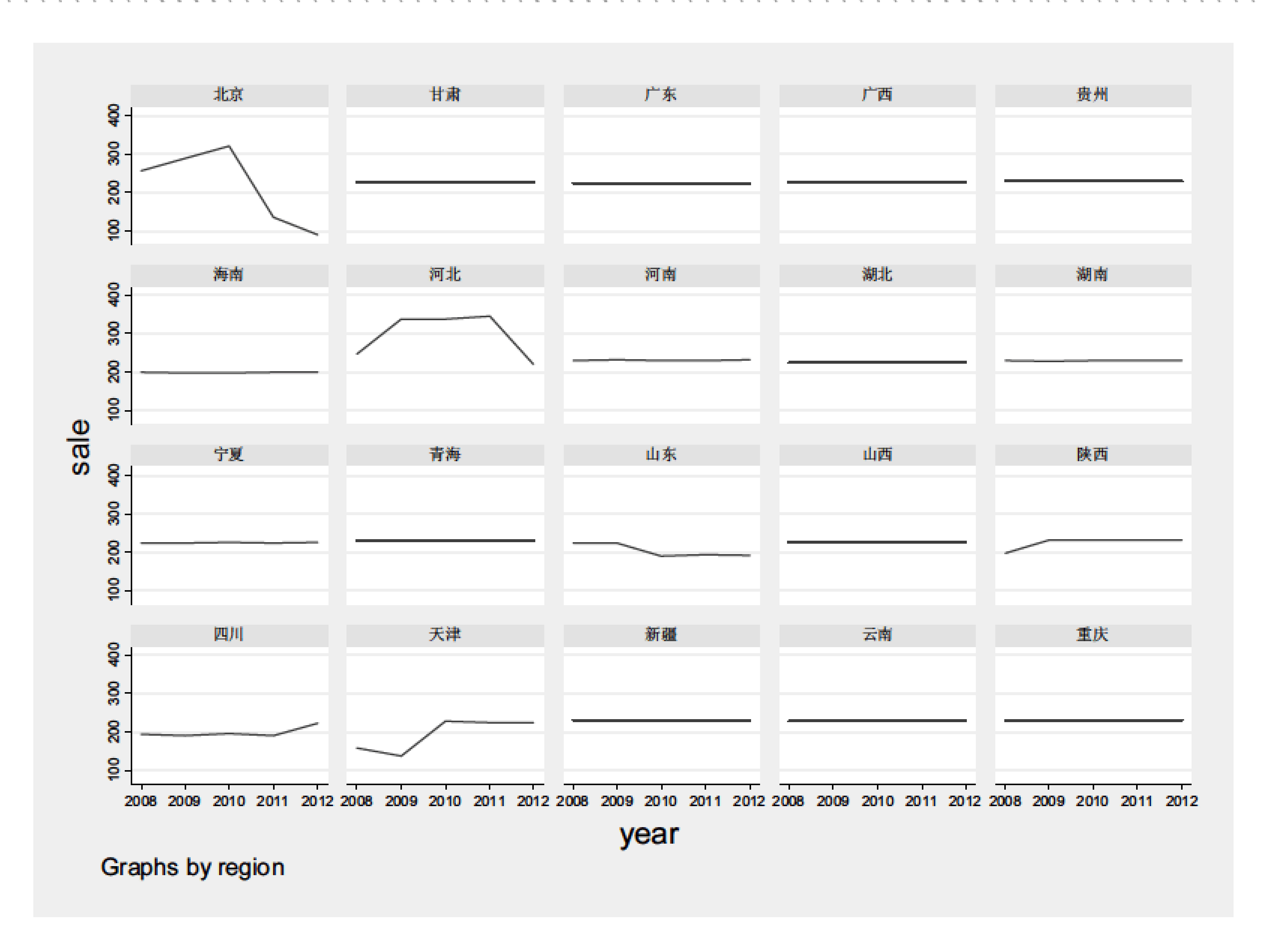


图 16.10　对每个个体显示“sale”变量的时间序列图

图 16.11 是对每个个体显示“cost”变量的时间序列图的结果。

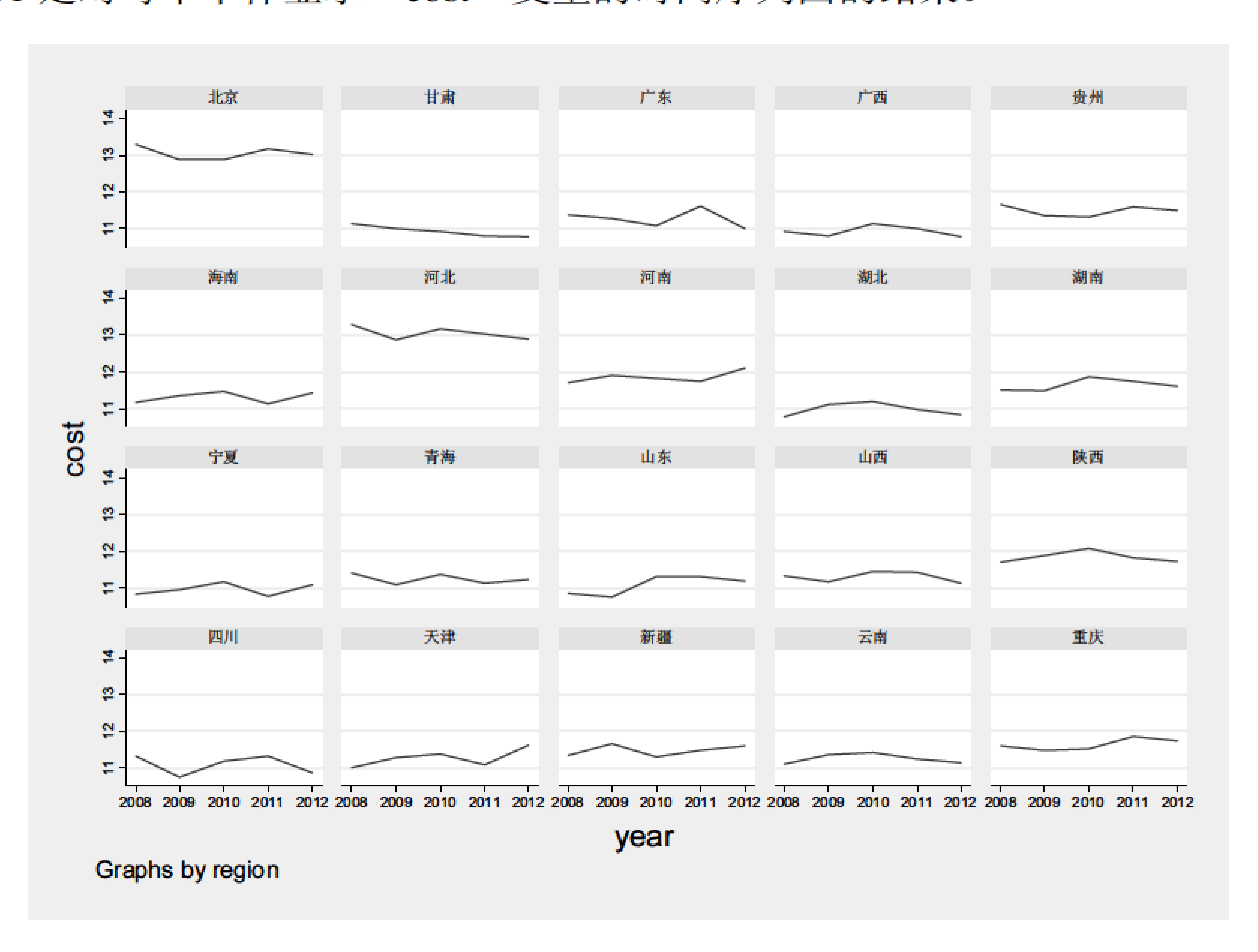


图 16.11　对每个个体显示“cost”变量的时间序列图

从图 16.11 可以看出，不同地区的促销成本的时间趋势是不一致的，有的地区变化非常平稳，有的地区先升后降，有的地区先降后升。

图 16.12 是对每个个体显示“profit”变量的时间序列图的结果。

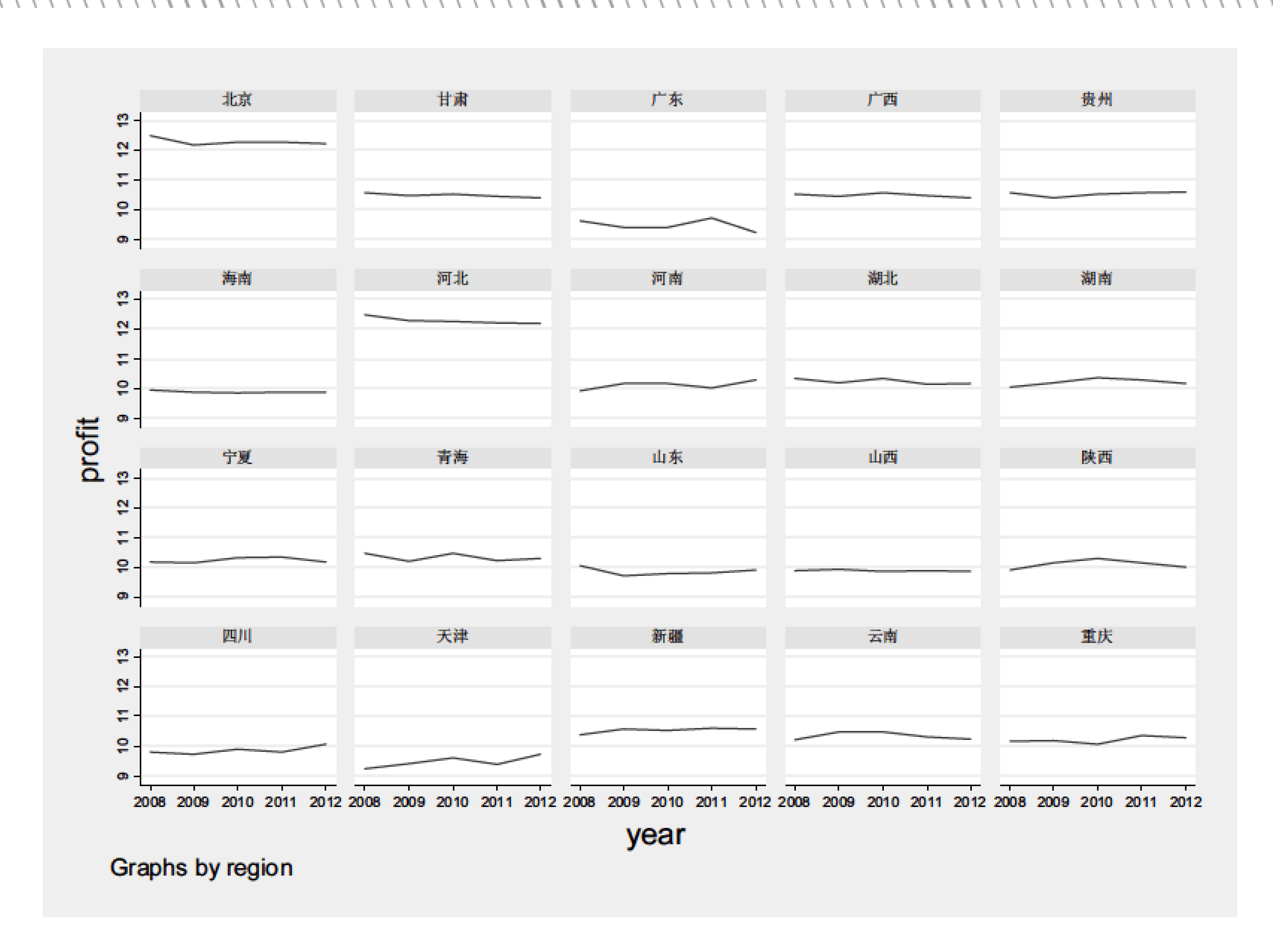


图 16.12　对每个个体显示“profit”变量的时间序列图

从图 16.12 可以看出，不同地区创造利润的时间趋势是不一致的，有的地区变化非常平稳，有的地区先升后降，有的地区先降后升。

图 16.13 是以 profit 为因变量，以 sale、cost 为自变量，进行最小二乘回归分析的结果。

```
. reg profit sale cost

      Source |       SS           df       MS      Number of obs   =       100
-------------+----------------------------------   F(  2,     97)  =     89.51
       Model |   33.828923         2  16.9144615   Prob > F        =    0.0000
    Residual |  18.3293904        97  .188962787   R-squared       =    0.6486
-------------+----------------------------------   Adj R-squared   =    0.6413
       Total |  52.1583134        99  .526851651   Root MSE        =     .4347

------------------------------------------------------------------------------
      profit |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        sale |   .0041186   .0014083     2.92   0.004     .0013235    .0069138
        cost |    .862813   .0755204    11.42   0.000     .7129259      1.0127
       _cons |  -.4981994    .823319    -0.61   0.547     -2.13226    1.135861
------------------------------------------------------------------------------
```

图 16.13　普通最小二乘回归分析

从上述分析结果中可以得到很多信息。可以看出共有 100 个样本参与了分析，模型的 F 值(2, 97) = 89.51，P 值（Prob > F）= 0.0000，说明模型整体上是很显著的。模型的可决系数（R-squared）为 0.6486，模型修正的可决系数（Adj R-squared）为 0.6413，说明模型的解释能力也是非常好的。

变量 sale 的系数标准误是 0.0014083，t 值为 2.92，P 值为 0.004，系数是非常显著的，95%的置信区间为[0.0013235, 0.0069138]。变量 cost 的系数标准误是 0.0755204，t 值为 11.42，P 值为 0.000，系数是非常显著的，95%的置信区间为[0.7129259, 1.0127]。常数项的系数标准误是 0.823319，t 值为-0.61，P 值为 0.547，系数是不显著的，95%的置信区间为[-2.13226,1.135861]。

从上述分析结果可以得到最小二乘模型的回归方程是：

profit = 0.0041186*sale+0.862813*cost -0.4981994

从上面的分析可以看出最小二乘线性模型的整体显著性、系数显著性以及模型的整体解释能力都很不错。得到的结论是该单位的创造利润情况与销售量和促销费用等都是显著呈正向变化的。

图 16.14 是以 profit 为因变量，以 sale、cost 为自变量，并使用以“region”为聚类变量的聚类稳健标准差，进行最小二乘回归分析的结果。

```
. reg profit sale cost,vce(cluster region)

Linear regression                               Number of obs =        100
                                                F(  2,     19) =      61.30
                                                Prob > F      =     0.0000
                                                R-squared     =     0.6486
                                                Root MSE      =      .4347

                                (Std. Err. adjusted for 20 clusters in region)
------------------------------------------------------------------------------
             |               Robust
      profit |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        sale |   .0041186   .0027939     1.47   0.157    -.0017291    .0099664
        cost |    .862813   .2199263     3.92   0.001      .402502    1.323124
       _cons |  -.4981994   1.986387    -0.25   0.805    -4.655755    3.659356
------------------------------------------------------------------------------
```

图 16.14　以“region”为聚类变量的聚类稳健标准差进行最小二乘回归分析

从图 16.14 中可以看出，使用以“region”为聚类变量的聚类稳健标准差进行最小二乘回归分析的结果与普通最小二乘回归分析得到的结果类似，只是 sale 变量系数的显著性有所下降。

图 16.15 是以 profit 为因变量，以 sale、cost 为自变量，并使用以“region”为聚类变量的聚类稳健标准差，进行固定效应回归分析的结果。

```
. xtreg profit sale cost,fe vce(cluster region)

Fixed-effects (within) regression               Number of obs      =       100
Group variable: region                          Number of groups   =        20

R-sq:  within  = 0.3637                         Obs per group: min =         5
       between = 0.6619                                        avg =       5.0
       overall = 0.6397                                        max =         5

                                                F(2,19)            =     10.92
corr(u_i, Xb)  = 0.6171                         Prob > F           =    0.0007

                                (Std. Err. adjusted for 20 clusters in region)
------------------------------------------------------------------------------
             |               Robust
      profit |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        sale |   .0008134    .000416     1.96   0.065    -.0000573     .001684
        cost |   .3855897   .0985735     3.91   0.001      .179273    .5919063
       _cons |   5.725855   1.122047     5.10   0.000     3.377383    8.074326
-------------+----------------------------------------------------------------
     sigma_u |  .55435378
     sigma_e |  .09590366
         rho |  .97094045   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

图 16.15　进行固定效应回归分析

从图 16.15 中可以看到共有 20 组，每组 5 个，共有 100 个样本参与了固定效应回归分析。模型的 F 值是 10.92，显著性 P 值为 0.0007，模型是非常显著的。模型组内 R 方是 0.3637（within = 0.3637），说明单位内解释的变化比例是 36.37%。模型组间 R 方是 0.6619（between = 0.6619），说明单位间解释的变化比例是 66.19%。模型总体 R 方是 0.6397（overall = 0.6397），说明总的解释变化比例是 63.97%。模型的解释能力还是可以接受的。观察模型中各个变量系数的显著性 P 值，发现也都是比较显著的。此外，观察图 16.15 中的最后一行，rho=0.97094045，说明复合扰动项的方差主要来自个体效应而不是时间效应的变动，这一点在后面的分析中也可以得到验证。

图 16.16 是以 profit 为因变量，以 sale、cost 为自变量，进行固定效应回归分析的结果。

```
. xtreg profit sale cost,fe

Fixed-effects (within) regression               Number of obs      =       100
Group variable: region                          Number of groups   =        20

R-sq:  within  = 0.3637                         Obs per group: min =         5
       between = 0.6619                                        avg =       5.0
       overall = 0.6397                                        max =         5

                                                F(2,78)            =     22.30
corr(u_i, Xb)  = 0.6171                         Prob > F           =    0.0000

------------------------------------------------------------------------------
      profit |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        sale |   .0008134   .0003772     2.16   0.034     .0000625    .0015643
        cost |   .3855897   .0596713     6.46   0.000     .2667932    .5043862
       _cons |   5.725855    .696736     8.22   0.000      4.33876    7.112949
-------------+----------------------------------------------------------------
     sigma_u |  .55435378
     sigma_e |  .09590366
         rho |  .97094045   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(19, 78) =   100.78              Prob > F = 0.0000
```

图 16.16　普通固定效应回归分析

本结果相对于使用以“region”为聚类变量的聚类稳健标准差进行固定效应回归分析的结果在变量系数显著性上有所提高。此外，在图 16.16 的最下面一行，可以看到“（F test that all u_i=0:　F(19, 78) = 100.78　Prob > F = 0.0000）”显著拒绝了所有各个样本没有自己的截距项的原假设，所以我们可以初步认为每个个体用于与众不同的截距项，也就是说固定效应模型是在一定程度上优于普通最小二乘回归模型的。这一点也在后续的深入分析中得到了验证。

图 16.17 存储的是固定效应回归分析估计结果。选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 16.17 所示的变量_est_fe 的相关数据。

	sale	cost	profit	diqu	region	year1	year2	year3	year4	year5	_est_fe
1	256	13.28039	12.47652	北京	北京	1	0	0	0	0	1
2	289	12.88284	12.1826	北京	北京	0	1	0	0	0	1
3	321	12.86566	12.26754	北京	北京	0	0	1	0	0	1
4	135	13.166	12.25672	北京	北京	0	0	0	1	0	1
5	89	13.01277	12.21607	北京	北京	0	0	0	0	1	1
6	226.4084	11.14041	10.55451	甘肃	甘肃	1	0	0	0	0	1
7	226.3114	11.0021	10.4631	甘肃	甘肃	0	1	0	0	0	1
8	226.2307	10.91509	10.51732	甘肃	甘肃	0	0	1	0	0	1
9	226.1334	10.80771	10.43588	甘肃	甘肃	0	0	0	1	0	1
10	226.0475	10.77687	10.39666	甘肃	甘肃	0	0	0	0	1	1
11	224.4039	11.38623	9.612467	广东	广东	1	0	0	0	0	1
12	224.2034	11.28099	9.401787	广东	广东	0	1	0	0	0	1
13	223.8923	11.07906	9.388487	广东	广东	0	0	1	0	0	1
14	224.4146	11.61005	9.714746	广东	广东	0	0	0	1	0	1
15	223.5251	11.00874	9.236008	广东	广东	0	0	0	0	1	1
16	226.2307	10.91509	10.51732	广西	广西	1	0	0	0	0	1
17	226.1334	10.80771	10.43588	广西	广西	0	1	0	0	0	1
18	226.4084	11.14041	10.55451	广西	广西	0	0	1	0	0	1
19	226.3114	11.0021	10.4631	广西	广西	0	0	0	1	0	1
20	226.0475	10.77687	10.39666	广西	广西	0	0	0	0	1	1
21	230.4395	11.65529	10.57132	贵州	贵州	1	0	0	0	0	1
22	230.2526	11.35158	10.38807	贵州	贵州	0	1	0	0	0	1
23	230.1745	11.30836	10.52889	贵州	贵州	0	0	1	0	0	1
24	230.4235	11.59451	10.56721	贵州	贵州	0	0	0	1	0	1
25	230.3779	11.48555	10.59037	贵州	贵州	0	0	0	0	1	1
26	201	11.17325	9.934502	海南	海南	1	0	0	0	0	1
27	198	11.33976	9.873029	海南	海南	0	1	0	0	0	1
28	199	11.46163	9.853772	海南	海南	0	0	1	0	0	1
29	201	11.12873	9.864227	海南	海南	0	0	0	1	0	1

图 16.17　固定效应回归分析估计结果

图 16.18 是构建最小二乘虚拟变量模型来分析固定效应模型是否优于最小二乘回归分析的分析结果。

```
. xi:xtreg  profit sale cost  i.region,vce(cluster region)
i.region          _Iregion_1-20       (naturally coded; _Iregion_1 omitted)

Random-effects GLS regression                   Number of obs      =       100
Group variable: region                          Number of groups   =        20

R-sq:  within  = 0.3637                         Obs per group: min =         5
       between = 1.0000                                        avg =       5.0
       overall = 0.9862                                        max =         5

                                                Wald chi2(2)       =         .
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =         .

                                   (Std. Err. adjusted for 20 clusters in region)

             |               Robust
      profit |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        sale |   .0008134   .0004639     1.75   0.080    -.0000958    .0017226
        cost |   .3855897   .1099256     3.51   0.000     .1701395    .6010398
  _Iregion_2 |  -.9982993   .2328748    -4.29   0.000    -1.454726    -.541873
  _Iregion_3 |  -2.132222   .1948403   -10.94   0.000    -2.514102   -1.750342
  _Iregion_4 |  -.9982993   .2328748    -4.29   0.000    -1.454726    -.541873
  _Iregion_5 |  -1.158279   .1726895    -6.71   0.000    -1.496744   -.8198136
  _Iregion_6 |  -1.715056   .1897156    -9.04   0.000    -2.086892   -1.343221
  _Iregion_7 |  -.0653463   .0372679    -1.75   0.080      -.13839    .0076973
  _Iregion_8 |  -1.725314   .1320758   -13.06   0.000    -1.984178    -1.46645
  _Iregion_9 |  -1.257426    .227647    -5.52   0.000    -1.703606   -.8112463
 _Iregion_10 |  -1.541913   .1551301    -9.94   0.000    -1.845963   -1.237864
 _Iregion_11 |  -1.257426    .227647    -5.52   0.000    -1.703606   -.8112463
 _Iregion_12 |  -1.263272    .197374    -6.40   0.000    -1.650118   -.8764264
 _Iregion_13 |  -1.662056   .2135228    -7.78   0.000    -2.080553   -1.243559
 _Iregion_14 |  -1.736067   .1913307    -9.07   0.000    -2.111068   -1.361065
 _Iregion_15 |  -1.719944   .1315194   -13.08   0.000    -1.977717   -1.462171
 _Iregion_16 |  -1.657931   .2132589    -7.77   0.000    -2.075911   -1.239951
 _Iregion_17 |   -2.10845   .1931195   -10.92   0.000    -2.486957   -1.729942
 _Iregion_18 |  -1.158279   .1726895    -6.71   0.000    -1.496744   -.8198136
 _Iregion_19 |  -1.263272    .197374    -6.40   0.000    -1.650118   -.8764264
 _Iregion_20 |  -1.541913   .1551301    -9.94   0.000    -1.845963   -1.237864
       _cons |   7.073893   1.422255     4.97   0.000     4.286323    9.861462
-------------+----------------------------------------------------------------
     sigma_u |          0
     sigma_e |  .09590366
         rho |          0   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

图 16.18　构建最小二乘虚拟变量模型

从图 16.18 中可以看出，大多数个体虚拟变量的显著性 P 值都是小于 0.05 的，所以我们可以非常有把握地认为可以拒绝“所有个体的虚拟变量皆为 0”的原假设，也就是说固定效应模型是优于普通最小二乘回归模型的。

图 16.19 是创建年度变量的多个虚拟变量的结果。选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 16.19 所示的变量 year1~year5 的相关数据。

	sale	cost	profit	diqu	region	year1	year2	year3	year4	year5
1	256	13.28039	12.47652	北京	北京	1	0	0	0	0
2	289	12.88284	12.1826	北京	北京	0	1	0	0	0
3	321	12.86566	12.26754	北京	北京	0	0	1	0	0
4	135	13.166	12.25672	北京	北京	0	0	0	1	0
5	89	13.01277	12.21607	北京	北京	0	0	0	0	1
6	226.4084	11.14041	10.55451	甘肃	甘肃	1	0	0	0	0
7	226.3114	11.0021	10.4631	甘肃	甘肃	0	1	0	0	0
8	226.2307	10.91509	10.51732	甘肃	甘肃	0	0	1	0	0
9	226.1334	10.80771	10.43588	甘肃	甘肃	0	0	0	1	0
10	226.0475	10.77687	10.39666	甘肃	甘肃	0	0	0	0	1
11	224.4039	11.38623	9.612467	广东	广东	1	0	0	0	0
12	224.2034	11.28099	9.401787	广东	广东	0	1	0	0	0
13	223.8923	11.07906	9.388487	广东	广东	0	0	1	0	0
14	224.4146	11.61005	9.714746	广东	广东	0	0	0	1	0
15	223.5251	11.00874	9.236008	广东	广东	0	0	0	0	1
16	226.2307	10.91509	10.51732	广西	广西	1	0	0	0	0
17	226.1334	10.80771	10.43588	广西	广西	0	1	0	0	0
18	226.4084	11.14041	10.55451	广西	广西	0	0	1	0	0
19	226.3114	11.0021	10.4631	广西	广西	0	0	0	1	0
20	226.0475	10.77687	10.39666	广西	广西	0	0	0	0	1
21	230.4395	11.65529	10.57132	贵州	贵州	1	0	0	0	0
22	230.2526	11.35158	10.38807	贵州	贵州	0	1	0	0	0
23	230.1745	11.30836	10.52889	贵州	贵州	0	0	1	0	0
24	230.4235	11.59451	10.56721	贵州	贵州	0	0	0	1	0
25	230.3779	11.48555	10.59037	贵州	贵州	0	0	0	0	1
26	201	11.17325	9.934502	海南	海南	1	0	0	0	0
27	198	11.33976	9.873029	海南	海南	0	1	0	0	0
28	199	11.46163	9.853772	海南	海南	0	0	1	0	0

图 16.19　创建年度变量的多个虚拟变量

图 16.20 是构建双向固定效应模型的分析结果。

```
. xtreg profit sale cost year2-year5,fe vce(cluster region)

Fixed-effects (within) regression               Number of obs      =       100
Group variable: region                          Number of groups   =        20

R-sq:  within  = 0.3714                         Obs per group: min =         5
       between = 0.6628                                        avg =       5.0
       overall = 0.6397                                        max =         5

                                                F(6,19)            =      6.27
corr(u_i, Xb)  = 0.6203                         Prob > F           =    0.0009

                              (Std. Err. adjusted for 20 clusters in region)
------------------------------------------------------------------------------
             |               Robust
      profit |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        sale |    .000841   .0004133     2.04   0.056     -.000024     .001706
        cost |   .3796737   .1023562     3.71   0.001     .1654398    .5939076
       year2 |  -.0227204   .0365359    -0.62   0.541     -.099191    .0537502
       year3 |  -.0020958   .0370119    -0.06   0.955    -.0795625     .075371
       year4 |   -.013553    .035162    -0.39   0.704    -.0871479    .0600418
       year5 |   .0018696   .0390425     0.05   0.962    -.0798473    .0835864
       _cons |   5.794876   1.163568     4.98   0.000       3.3595    8.230251
-------------+----------------------------------------------------------------
     sigma_u |  .55623368
     sigma_e |  .09786431
         rho |  .96997422   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

图 16.20　构建双向固定效应模型

从图 16.20 中可以看出，全部 year 虚拟变量的显著性 P 值都是远大于 0.05 的，所以我们可以初步认为模型中不应包含时间效应。值得说明的是，在构建双向固定效应模型时并没有把 year1 列入进去，这是因为 year1 被视为基期，也就是模型中的常数项。

图 16.21 是在上步回归的基础上，通过测试各虚拟变量的系数联合显著性来检验是否应该在模型中纳入时间效应的检验结果。

```
. test year2 year3 year4 year5

 ( 1)  year2 = 0
 ( 2)  year3 = 0
 ( 3)  year4 = 0
 ( 4)  year5 = 0

       F(  4,    19) =    0.30
            Prob > F =    0.8774
```

图 16.21　测试各虚拟变量系数联合显著性

从图 16.21 中可以看出，各变量系数的联合显著性是非常差的，接受了没有时间效应的初始假设，所以我们进一步验证了模型中不必包含时间效应项的结论。

图 16.22 是以 profit 为因变量，以 sale、cost 为自变量，并使用以“region”为聚类变量的聚类稳健标准差，进行随机效应回归分析的结果。

从图 16.22 可以看出，随机效应回归分析的结果与固定效应回归分析的结果大同小异，只是部分变量的显著性水平得到了进一步的提高。

图 16.23 是在上步回归的基础上，进行假设检验来判断随机效应模型是否优于最小二乘回归模型的结果。

从图 16.23 可以看出，假设检验非常显著地拒绝了不存在个体随机效应的原假设，也就是说，随机效应模型是在一定程度上优于普通最小二乘回归分析模型的。

```
. xtreg profit sale cost,re vce(cluster region)

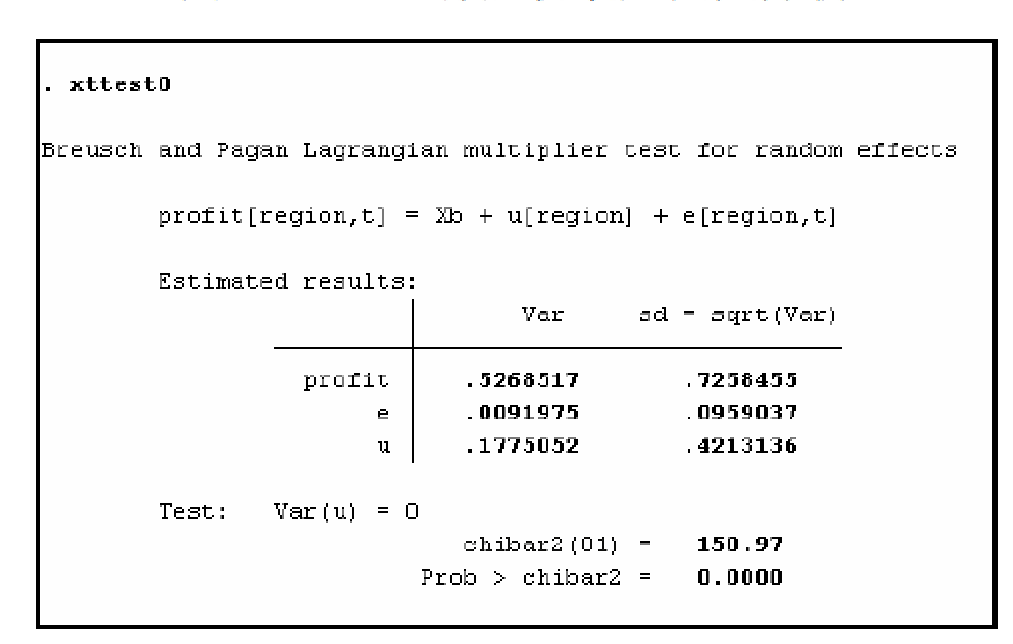
Random-effects GLS regression                   Number of obs      =       100
Group variable: region                          Number of groups   =        20

R-sq:  within  = 0.3637                         Obs per group: min =         5
       between = 0.6615                                        avg =       5.0
       overall = 0.6394                                        max =         5

                                                Wald chi2(2)       =     57.98
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000

                              (Std. Err. adjusted for 20 clusters in region)
------------------------------------------------------------------------------
             |               Robust
      profit |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        sale |    .000941   .0004111     2.29   0.022     .0001354    .0017467
        cost |   .4552322   .1038988     4.38   0.000     .2515942    .6588701
       _cons |   4.897379   1.115396     4.39   0.000     2.711243    7.083515
-------------+----------------------------------------------------------------
     sigma_u |  .42131364
     sigma_e |  .09590366
         rho |  .95073713   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

图 16.22　进行随机效应回归分析



```
. xttest0

Breusch and Pagan Lagrangian multiplier test for random effects

        profit[region,t] = Xb + u[region] + e[region,t]

        Estimated results:
                         |       Var     sd = sqrt(Var)
                ---------+-----------------------------
                  profit |   .5268517       .7258455
                       e |   .0091975       .0959037
                       u |   .1775052       .4213136

        Test:   Var(u) = 0
                             chibar2(01) =   150.97
                          Prob > chibar2 =   0.0000
```

图 16.23　进行假设检验

图 16.24 是以 profit 为因变量，以 sale、cost 为自变量，并使用最大似然估计方法进行随机效应回归分析的结果。

```
. xtreg profit sale cost,mle

Fitting constant-only model:
Iteration 0:   log likelihood = -.34409155
Iteration 1:   log likelihood =  17.020843
Iteration 2:   log likelihood =  18.904008
Iteration 3:   log likelihood =  19.202858
Iteration 4:   log likelihood =  19.218547
Iteration 5:   log likelihood =  19.218613

Fitting full model:
Iteration 0:   log likelihood =  7.9773037
Iteration 1:   log likelihood =  19.164908
Iteration 2:   log likelihood =  38.281199
Iteration 3:   log likelihood =   42.70826
Iteration 4:   log likelihood =  43.214387
Iteration 5:   log likelihood =  43.225571
Iteration 6:   log likelihood =  43.225578

Random-effects ML regression                    Number of obs      =       100
Group variable: region                          Number of groups   =        20

Random effects u_i ~ Gaussian                   Obs per group: min =         5
                                                               avg =       5.0
                                                               max =         5

                                                LR chi2(2)         =     48.01
Log likelihood  =  43.225578                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
      profit |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        sale |   .0008985    .000374     2.40   0.016     .0001655    .0016315
        cost |   .4326386   .0588545     7.35   0.000      .317286    .5479913
       _cons |   5.166409   .6975167     7.41   0.000     3.799301    6.533516
-------------+----------------------------------------------------------------
    /sigma_u |   .5208324   .0855846                      .3774212    .7187365
    /sigma_e |    .095091    .007579                      .0813385    .1111686
         rho |   .9677417   .0115572                      .9376186    .9846948
------------------------------------------------------------------------------
Likelihood-ratio test of sigma_u=0: chibar2(01)=  200.57 Prob>=chibar2 = 0.000
```

图 16.24　使用最大似然估计方法进行随机效应回归分析

从图 16.24 可以看出，使用最大似然估计方法的随机效应回归分析的结果与使用以“region”为聚类变量的聚类稳健标准差的随机效应回归分析的结果大同小异，只是部分变量的显著性水平得到了进一步的提高。

图 16.25 是以 profit 为因变量，以 sale、cost 为自变量，并使用组间估计量，进行组间估计量回归分析的结果。

```
. xtreg profit sale cost,be

Between regression (regression on group means)  Number of obs      =       100
Group variable: region                          Number of groups   =        20

R-sq:  within  = 0.1532                         Obs per group: min =         5
       between = 0.7013                                        avg =       5.0
       overall = 0.5968                                        max =         5

                                                F(2,17)            =     19.95
sd(u_i + avg(e_i.))=  .4234911                  Prob > F           =    0.0000

------------------------------------------------------------------------------
      profit |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        sale |   .0104226   .0056309     1.85   0.082    -.0014576    .0223028
        cost |   .7736021   .1950808     3.97   0.001     .3620176    1.185187
       _cons |  -.8923599   1.857947    -0.48   0.637    -4.812285    3.027565
------------------------------------------------------------------------------
```

图 16.25　使用组间估计量进行组间估计量回归分析

从图 16.25 可以看出，使用组间估计量进行回归分析的结果较固定效应模型、随机效应模型在模型的解释能力以及变量系数的显著性上都有所降低。

16.1.5　案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

延伸：关于模型的选择问题

在前面的分析过程部分，我们使用各种分析方法对本节涉及的案例进行了详细具体的分析。读者们看到众多的分析方法时可能会有眼花缭乱的感觉，那么我们最终应该选择哪种分析方法来构建模型呢？答案当然是具体问题具体分析，然而我们也有统计方法和统计经验作为决策参考。例如，在本例中，已经证明了固定效应模型和随机效应模型都要好于普通最小二乘回归模型。而对于组间估计量模型来说，它通常用于数据质量不好的时候，而且会损失较多的信息，所以很多时候我们仅仅将其作为一种对照的估计方法。那么剩下的问题就是选择固定效应模型还是随机效应模型的问题。在前面分析的基础上，操作命令如下。

- xtreg profit sale cost,re：本命令的含义是以 profit 为因变量，以 sale、cost 为自变量，进行随机效应回归分析。
- estimates store re：本命令的含义是存储随机效应回归分析的估计结果。
- hausman fe re,constant sigmamore：本命令的含义是进行豪斯曼检验，并据此判断应该选择固定效应模型还是随机效应模型。

在命令窗口输入命令并按回车键进行确认，结果如图 16.26~图 16.28 所示。

图 16.26 是以 profit 为因变量，以 sale、cost 为自变量，进行随机效应回归分析的结果。

```
. xtreg profit sale cost,re

Random-effects GLS regression                   Number of obs      =       100
Group variable: region                          Number of groups   =        20

R-sq:  within  = 0.3637                         Obs per group: min =         5
       between = 0.6615                                        avg =       5.0
       overall = 0.6394                                        max =         5

                                                Wald chi2(2)       =     62.84
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
      profit |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        sale |    .000941   .0003979     2.37   0.018     .0001612    .0017209
        cost |   .4552322   .0592611     7.68   0.000     .3390826    .5713817
       _cons |   4.897379   .6983754     7.01   0.000     3.528588    6.266169
-------------+----------------------------------------------------------------
     sigma_u |  .42131364
     sigma_e |  .09590366
         rho |  .95073713   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

图 16.26 进行随机效应回归分析

对该回归分析结果的详细解读我们在前面也多次讲述，此次不再重复讲解。

图 16.27 存储的是随机效应回归分析估计结果。选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 16.27 所示的变量_est_re 的相关数据。

	sale	cost	profit	diqu	region	year1	year2	year3	year4	year5	_est_fe	_est_re
1	256	13.28039	12.47652	北京	北京	1	0	0	0	0	1	1
2	289	12.88284	12.1826	北京	北京	0	1	0	0	0	1	1
3	321	12.86566	12.26754	北京	北京	0	0	1	0	0	1	1
4	135	13.166	12.25672	北京	北京	0	0	0	1	0	1	1
5	89	13.01277	12.21607	北京	北京	0	0	0	0	1	1	1
6	226.4084	11.14041	10.55451	甘肃	甘肃	1	0	0	0	0	1	1
7	226.3114	11.0021	10.4631	甘肃	甘肃	0	1	0	0	0	1	1
8	226.2307	10.91509	10.51732	甘肃	甘肃	0	0	1	0	0	1	1
9	226.1334	10.80771	10.43588	甘肃	甘肃	0	0	0	1	0	1	1
10	226.0475	10.77687	10.39666	甘肃	甘肃	0	0	0	0	1	1	1
11	224.4039	11.38623	9.612467	广东	广东	1	0	0	0	0	1	1
12	224.2034	11.28099	9.401787	广东	广东	0	1	0	0	0	1	1
13	223.8923	11.07906	9.388487	广东	广东	0	0	1	0	0	1	1
14	224.4146	11.61005	9.714746	广东	广东	0	0	0	1	0	1	1
15	223.5251	11.00874	9.236008	广东	广东	0	0	0	0	1	1	1
16	226.2307	10.91509	10.51732	广西	广西	1	0	0	0	0	1	1
17	226.1334	10.80771	10.43588	广西	广西	0	1	0	0	0	1	1
18	226.4084	11.14041	10.55451	广西	广西	0	0	1	0	0	1	1
19	226.3114	11.0021	10.4631	广西	广西	0	0	0	1	0	1	1
20	226.0475	10.77687	10.39666	广西	广西	0	0	0	0	1	1	1
21	230.4395	11.65529	10.57132	贵州	贵州	1	0	0	0	0	1	1
22	230.2526	11.35158	10.38807	贵州	贵州	0	1	0	0	0	1	1
23	230.1745	11.30836	10.52889	贵州	贵州	0	0	1	0	0	1	1
24	230.4235	11.59451	10.56721	贵州	贵州	0	0	0	1	0	1	1
25	230.3779	11.48555	10.59037	贵州	贵州	0	0	0	0	1	1	1
26	201	11.17325	9.934502	海南	海南	1	0	0	0	0	1	1
27	198	11.33976	9.873029	海南	海南	0	1	0	0	0	1	1
28	199	11.46163	9.853772	海南	海南	0	0	1	0	0	1	1

图 16.27 查看数据

图 16.28 是进行豪斯曼检验的结果。

豪斯曼检验的原假设是使用随机效应模型。图 16.28 中显示的显著性 P 值（Prob>chi2 =0.0061）远远低于 5%，所以我们拒绝初始假设，认为使用固定效应模型是更为合理的。

综上所述，我们应该构建固定效应模型来描述变量之间的回归关系。

```
. hausman fe re,constant sigmamore

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |      fe           re          Difference          S.E.
-------------+----------------------------------------------------------------
        sale |   .0008134      .000941       -.0001277         .000038
        cost |   .3855897     .4552322       -.0696425        .0220623
       _cons |   5.725855     4.897379        .8284759        .2396264
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from xtreg
            B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(3) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =       12.40
                Prob>chi2 =      0.0061
                (V_b-V_B is not positive definite)
```

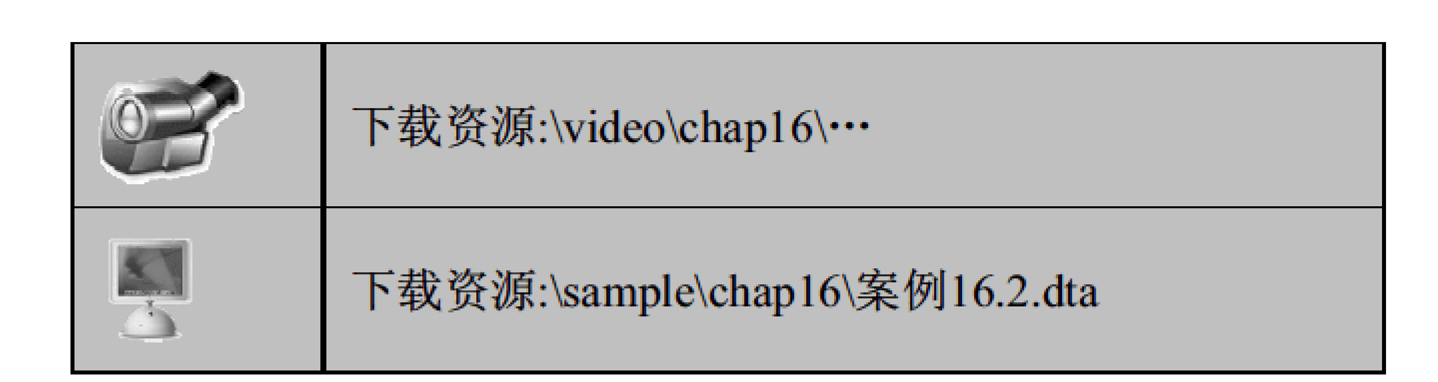
图 16.28　进行豪斯曼检验

16.2　实例二——长面板数据分析

16.2.1　长面板数据分析的功能与意义

长面板数据是面板数据的一种，其主要特征是时间维度比较大而横截面维度相对较小，或者说，同一期间内被观测的期间较多而被观测的个体数量较少。长面板数据分析相对而言更加关注设定扰动项相关的具体形式，一般使用可行广义最小二乘法进行估计。这又分为两种情形：一种是仅解决组内自相关的可行广义最小二乘估计；另一种是同时处理组内自相关与组间同期相关的可行广义最小二乘估计。下面就以实例的方式来介绍一下这几种方法的具体应用。

16.2.2　相关数据来源



	下载资源:\video\chap16\…
	下载资源:\sample\chap16\案例16.2.dta

【例 16.2】B 公司是一家保险公司，经营范围遍布全国 10 个省市，各省市连锁店 2001—2010 年的相关经营数据包括保费收入、赔偿支出以及创造利润等，如表 16.2 所示。试用多种长面板数据回归分析方法深入研究保费收入、赔偿支出对创造利润的影响关系。

表 16.2　B 公司各省市保费收入、赔偿支出以及创造利润数据（2001—2010 年）

年份	保费收入/万元	赔偿支出/万元	创造利润/万元	省市
2001	259.587	58.56	26.211	北京
2002	261.083	52.23	21.039	北京
2003	259.296	44.81	20.201	北京
2004	257.546	39.35	19.536	北京
2005	255.723	38.68	21.268	北京

（续表）

年份	保费收入/万元	赔偿支出/万元	创造利润/万元	省市
2006	29.865	9.5	1.903	北京
…	…	…	…	…
2005	23.154	6.04	1.026	浙江
2006	30.892	6.89	3.835	浙江
2007	30.594	6	3.5	浙江
2008	30.348	5.5	3.695	浙江
2009	30.054	4.94	3.406	浙江
2010	29.797	4.79	3.275	浙江

16.2.3 Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有 5 个变量，分别是年份、保费收入、赔偿支出、创造利润以及省市。我们把年份变量定义为 year，把保费收入变量定义为 income，把赔偿支出变量定义为 cost，把创造利润变量定义为 profit，把省市变量定义为 shengshi。变量类型及长度采取系统默认方式，然后录入相关数据。相关操作在第 1 章中已有详细讲述。录入完成后数据如图 16.29 所示。

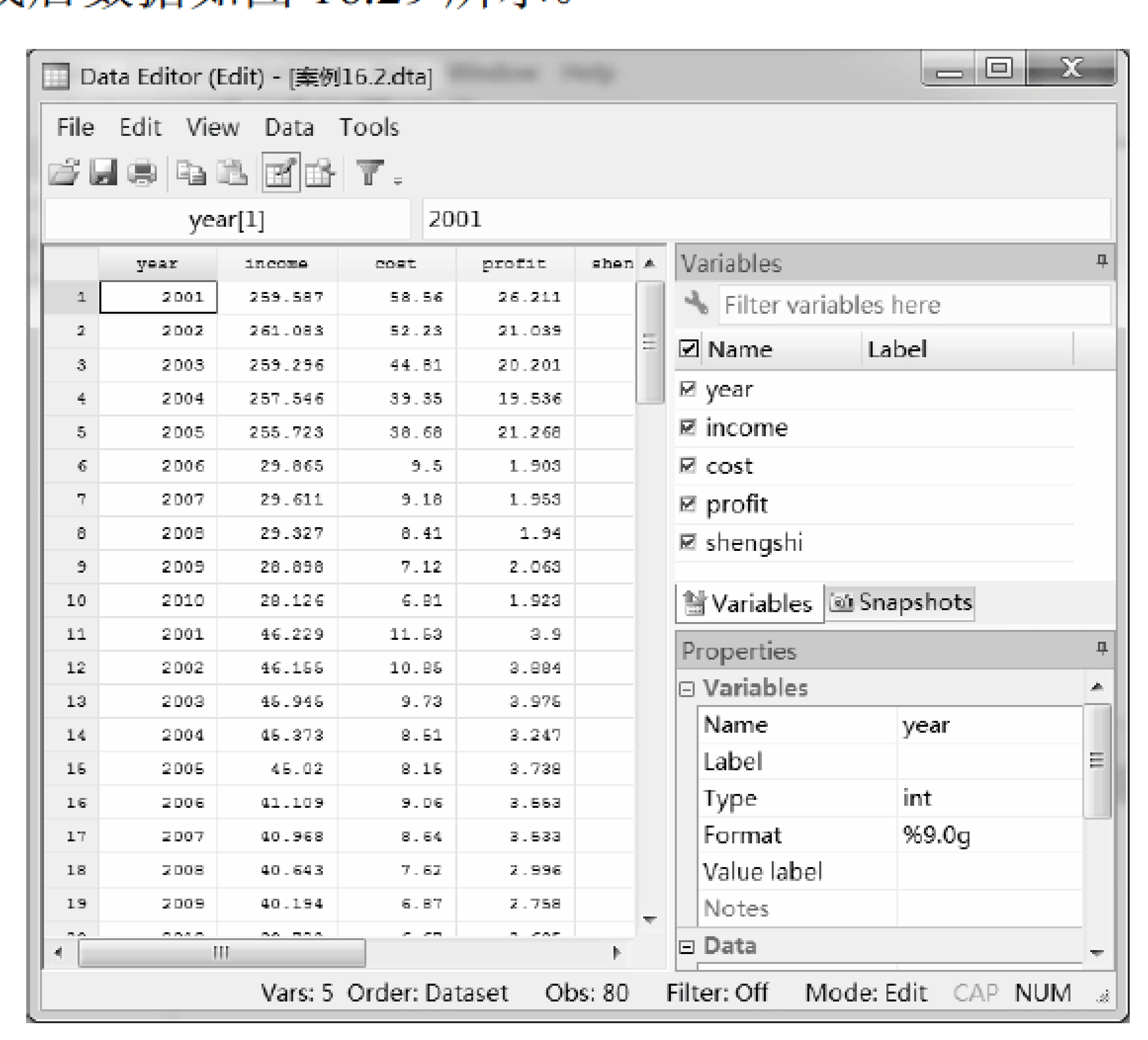


图 16.29　案例 16.2 数据

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入如下命令。

- list year income cost profit：本命令的含义是对 4 个变量所包含的样本数据进行一一展示，以便简单直观地观测出数据的具体特征，为深入分析做好必要准备。
- encode shengshi,gen(region)：因为面板数据要求其中的个体变量取值必须为整数而且不允许有重复，所以我们需要对各个观测样本进行有序编号。本命令旨在将 shengshi 这一字符串变量转化为数值型变量，以便进行下一步操作。

- xtset region year：本命令的含义是对面板数据进行定义，其中横截面维度变量为我们上步生成的 region，时间序列变量为 year。
- xtdes：本命令旨在观测面板数据的结构，考察面板数据特征，为后续分析做好必要准备。
- xtsum：本命令旨在显示面板数据组内、组间以及整体的统计指标。
- xttab income：本命令旨在显示“income”变量组内、组间以及整体的分布频率。
- xttab cost：本命令旨在显示“cost”变量组内、组间以及整体的分布频率。
- xttab profit：本命令旨在显示“profit”变量组内、组间以及整体的分布频率。
- xtline income：本命令旨在对每个个体显示“income”变量的时间序列图。
- xtline cost：本命令旨在对每个个体显示“cost”变量的时间序列图。
- xtline profit：本命令旨在对每个个体显示“profit”变量的时间序列图。
- tab region,gen(region)：本命令旨在创建省市变量的多个虚拟变量。
- reg profit income cost region2-region8 year,vce(cluster region)：本命令的含义是以 profit 为因变量，以 income、cost 以及生成的各个地区虚拟变量为自变量，并使用以“region”为聚类变量的聚类稳健标准差，进行最小二乘回归分析。
- estimates store ols：本命令的含义是存储最小二乘回归分析的估计结果。
- xtpcse profit income cost region2-region8 year,corr(ar1)：本命令的含义是在仅考虑存在组内自相关，并且各组的自回归系数相同的情形下，以 profit 为因变量，以 income、cost 以及生成的各个地区虚拟变量为自变量，进行可行广义最小二乘回归分析。
- estimates store ar1：本命令的含义是存储上步可行广义最小二乘回归分析的估计结果。
- xtpcse profit income cost region2-region8 year,corr(psar1)：本命令的含义是在仅考虑存在组内自相关，并且各组的自回归系数不相同的情形下，以 profit 为因变量，以 income、cost 以及生成的各个地区虚拟变量为自变量，进行可行广义最小二乘回归分析。
- estimates store psar1：本命令的含义是存储上步可行广义最小二乘回归分析的估计结果。
- xtpcse profit income cost region2-region8 year,hetonly：本命令的含义是在不考虑存在自相关，仅考虑不同个体扰动项存在异方差的情形下，以 profit 为因变量，以 income、cost 以及生成的各个地区虚拟变量为自变量，进行可行广义最小二乘回归分析。
- estimates store hetonly：本命令的含义是存储上步可行广义最小二乘回归分析的估计结果。
- estimates table ols ar1 psar1 hetonly,b se：本命令的含义是展示将以上各种方法的系数估计值及标准差列表放到一起进行比较的结果。
- xtgls profit income cost region2-region8 year,panels(cor) cor(ar1)：本命令的含义是在假定不同个体的扰动项相互独立且有不同的方差，并且各组的自回归系数相同的情形下，以 profit 为因变量，以 income、cost 以及生成的各个地区虚拟变量为自变量，进行可行广义最小二乘回归分析。
- xtgls profit income cost region2-region8 year,panels(cor) cor(psar1)：本命令的含义是在假定不同个体的扰动项相互独立且有不同的方差，并且各组的自回归系数不相同的情形下，以 profit 为因变量，以 income、cost 以及生成的各个地区虚拟变量为自变量，进行可行广义最小二乘回归分析。

03 设置完毕后，按键盘上的回车键，等待输出结果。

16.2.4 结果分析

在 Stata 14.0 主界面的结果窗口可以看到如图 16.30~图 16.52 所示的分析结果。

图 16.30 是对数据进行展示的结果。它的目的是通过对变量所包含的样本数据进行一一展示，以便简单直观地观测出数据的具体特征，为深入分析做好必要准备。

```
. list year income cost profit
```

	year	income	cost	profit
1.	2001	259.587	58.56	26.211
2.	2002	261.083	52.23	21.039
3.	2003	259.296	44.81	20.201
4.	2004	257.546	39.35	19.536
5.	2005	255.723	38.68	21.268
6.	2006	29.865	9.5	1.903
7.	2007	29.611	9.18	1.953
8.	2008	29.327	8.41	1.94
9.	2009	28.898	7.12	2.063
10.	2010	28.126	6.81	1.923
11.	2001	46.229	11.53	3.9
12.	2002	46.155	10.85	3.884
13.	2003	45.945	9.73	3.975
14.	2004	45.373	8.51	3.247
15.	2005	45.02	8.15	3.738
16.	2006	41.109	9.06	3.553
17.	2007	40.968	8.64	3.533
18.	2008	40.643	7.62	2.996
19.	2009	40.194	6.87	2.758
20.	2010	39.722	6.67	2.685
21.	2001	44.038	14.15	3.148
22.	2002	44.017	12.49	2.933
23.	2003	43.513	10.95	2.575
24.	2004	42.88	9.99	2.322
25.	2005	42.122	9.69	2.638
26.	2006	52.523	17.85	2.936
27.	2007	51.976	14.67	2.582
28.	2008	51.144	13.62	2.579
29.	2009	50.047	12.53	2.226
30.	2010	48.943	12.05	2.023
31.	2001	24.495	8.27	1.779
32.	2002	24.408	8.25	1.811
33.	2003	24.083	7.26	1.992
34.	2004	23.478	5.22	2.346
35.	2005	22.774	4.7	1.665
36.	2006	26.116	7.18	3.042
37.	2007	26.102	6.67	2.634
38.	2008	25.75	5.8	2.531
39.	2009	25.464	5.09	2.61
40.	2010	25.203	4.8	3.108
41.	2001	25.308	11.02	1.656
42.	2002	25.281	8.81	1.495
43.	2003	24.779	7.93	1.211
44.	2004	24.02	6.48	1.195
45.	2005	23.154	6.04	1.026
46.	2006	30.892	6.89	3.835
47.	2007	30.594	6	3.5
48.	2008	30.348	5.5	3.695
49.	2009	30.054	4.94	3.406
50.	2010	29.797	4.79	3.275
51.	2001	259.587	58.56	26.211
52.	2002	261.083	52.23	21.039
53.	2003	259.296	44.81	20.201
54.	2004	257.546	39.35	19.536
55.	2005	255.723	38.68	21.268
56.	2006	29.865	9.5	1.903
57.	2007	29.611	9.18	1.953
58.	2008	29.327	8.41	1.94
59.	2009	28.898	7.12	2.063
60.	2010	28.126	6.81	1.923
61.	2001	24.495	8.27	1.779
62.	2002	24.408	8.25	1.811
63.	2003	24.083	7.26	1.992
64.	2004	23.478	5.22	2.346
65.	2005	22.774	4.7	1.665
66.	2006	26.116	7.18	3.042
67.	2007	26.102	6.67	2.634
68.	2008	25.75	5.8	2.531
69.	2009	25.464	5.09	2.61
70.	2010	25.203	4.8	3.108
71.	2001	25.308	11.02	1.656
72.	2002	25.281	8.81	1.495
73.	2003	24.779	7.93	1.211
74.	2004	24.02	6.48	1.195
75.	2005	23.154	6.04	1.026
76.	2006	30.892	6.89	3.835
77.	2007	30.594	6	3.5
78.	2008	30.348	5.5	3.695
79.	2009	30.054	4.94	3.406
80.	2010	29.797	4.79	3.275

图 16.30　展示数据

在如图 16.30 所示的分析结果中可以看出，数据的总体质量还是可以的，没有极端异常值，变量间的量纲差距也是可以接受的，可以进入下一步的分析。

图 16.31 是将 shengshi 这一字符串变量转化为数值型变量 region 的结果。选择“Data” | “Data Editor” | “Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 16.31 所示的变量 region 的相关数据。

	year	income	cost	profit	shengshi	region
1	2001	259.587	58.56	26.211	北京	北京
2	2002	261.083	52.23	21.039	北京	北京
3	2003	259.296	44.81	20.201	北京	北京
4	2004	257.546	39.35	19.536	北京	北京
5	2005	255.723	38.68	21.268	北京	北京
6	2006	29.865	9.5	1.903	北京	北京
7	2007	29.611	9.18	1.953	北京	北京
8	2008	29.327	8.41	1.94	北京	北京
9	2009	28.898	7.12	2.063	北京	北京
10	2010	28.126	6.81	1.923	北京	北京
11	2001	46.229	11.53	3.9	广东	广东
12	2002	46.155	10.85	3.884	广东	广东
13	2003	45.945	9.73	3.975	广东	广东
14	2004	45.373	8.51	3.247	广东	广东
15	2005	45.02	8.15	3.738	广东	广东
16	2006	41.109	9.06	3.553	广东	广东
17	2007	40.968	8.64	3.533	广东	广东
18	2008	40.643	7.62	2.996	广东	广东
19	2009	40.194	6.87	2.758	广东	广东
20	2010	39.722	6.67	2.685	广东	广东
21	2001	44.038	14.15	3.148	广西	广西
22	2002	44.017	12.49	2.933	广西	广西
23	2003	43.513	10.95	2.575	广西	广西
24	2004	42.88	9.99	2.322	广西	广西
25	2005	42.122	9.69	2.638	广西	广西
26	2006	52.523	17.85	2.936	广西	广西
27	2007	51.976	14.67	2.582	广西	广西
28	2008	51.144	13.62	2.579	广西	广西
29	2009	50.047	12.53	2.226	广西	广西

图 16.31　查看数据

图 16.32 是对面板数据进行定义的结果，其中横截面维度变量为上步生成的 region，时间序列变量为 year。

```
xtset region year
      panel variable:  region (strongly balanced)
       time variable:  year, 2001 to 2010
               delta:  1 unit
```

图 16.32　对面板数据进行定义

从图 16.32 可以看出这是一个平衡的面板数据。

图 16.33 是面板数据结构的结果。

```
. xtdes

  region:  1, 2, ..., 8                                    n =          8
    year:  2001, 2002, ..., 2010                           T =         10
           Delta(year) = 1 unit
           Span(year)  = 10 periods
           (region*year uniquely identifies each observation)

Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                        10      10      10        10        10      10      10

     Freq.  Percent    Cum. |  Pattern
 ---------------------------+------------
        8    100.00  100.00 |  1111111111
 ---------------------------+------------
        8    100.00         |  XXXXXXXXXX
```

图 16.33　面板数据结构

从图 16.33 可以看出该面板数据的横截面维度 region 为 1~8 共 8 个取值，时间序列维度 year 为 2001~2010 共 10 个取值，属于长面板数据，而且观测样本在时间上的分布也非常均匀。

图 16.34 是面板数据组内、组间以及整体的统计指标的结果。

```
. xtsum

Variable         |      Mean   Std. Dev.       Min        Max |    Observations
-----------------+--------------------------------------------+----------------
year     overall |    2005.5   2.890403       2001       2010 |     N =      80
         between |                    0     2005.5     2005.5 |     n =       8
         within  |             2.890403       2001       2010 |     T =      10
                 |                                            |
income   overall |  60.31106   75.89957     22.774    261.083 |     N =      80
         between |             52.28008    24.7873   143.9062 |     n =       8
         within  |             57.78336  -55.46914   177.4879 |     T =      10
                 |                                            |
cost     overall |   12.8525   13.41096        4.7      58.56 |     N =      80
         between |              9.26838      6.324     27.465 |     n =       8
         within  |             10.18515    -7.8025    43.9475 |     T =      10
                 |                                            |
profit   overall |  4.899112   6.471817      1.026     26.211 |     N =      80
         between |              4.27608     2.3518    11.8037 |     n =       8
         within  |             5.067804  -5.001587   19.30641 |     T =      10
                 |                                            |
shengshi overall |         .          .          .          . |     N =       0
         between |                    .          .          . |     n =       0
         within  |                    .          .          . |     T =       .
                 |                                            |
region   overall |       4.5   2.305744          1          8 |     N =      80
         between |              2.44949          1          8 |     n =       8
         within  |                    0        4.5        4.5 |     T =      10
```

图 16.34　板数据组内、组间以及整体的统计指标

在短面板数据中，同一时间段内的不同观测样本构成一个组。从图 16.34 中可以看出，变量 year 的组间标准差是 0，因为不同组的这一变量取值完全相同，同时变量 region 的组内标

准差也为 0，因为分布在同一组的数据属于同一个地区。

图 16.35 是“income”变量组内、组间以及整体的分布频率的结果。

. xttab income

income	Overall Freq.	Overall Percent	Between Freq.	Between Percent	Within Percent
22.774	2	2.50	2	25.00	10.00
23.154	2	2.50	2	25.00	10.00
23.478	2	2.50	2	25.00	10.00
24.02	2	2.50	2	25.00	10.00
24.083	2	2.50	2	25.00	10.00
24.408	2	2.50	2	25.00	10.00
24.495	2	2.50	2	25.00	10.00
24.779	2	2.50	2	25.00	10.00
25.203	2	2.50	2	25.00	10.00
25.281	2	2.50	2	25.00	10.00
25.308	2	2.50	2	25.00	10.00
25.464	2	2.50	2	25.00	10.00
25.75	2	2.50	2	25.00	10.00
26.102	2	2.50	2	25.00	10.00
26.116	2	2.50	2	25.00	10.00
28.126	2	2.50	2	25.00	10.00
28.898	2	2.50	2	25.00	10.00
29.327	2	2.50	2	25.00	10.00
29.611	2	2.50	2	25.00	10.00
29.797	2	2.50	2	25.00	10.00
29.865	2	2.50	2	25.00	10.00
30.054	2	2.50	2	25.00	10.00
30.348	2	2.50	2	25.00	10.00
30.594	2	2.50	2	25.00	10.00
30.892	2	2.50	2	25.00	10.00
39.722	1	1.25	1	12.50	10.00
40.194	1	1.25	1	12.50	10.00
40.643	1	1.25	1	12.50	10.00
40.968	1	1.25	1	12.50	10.00
41.109	1	1.25	1	12.50	10.00
42.122	1	1.25	1	12.50	10.00
42.88	1	1.25	1	12.50	10.00
43.513	1	1.25	1	12.50	10.00
44.017	1	1.25	1	12.50	10.00
44.038	1	1.25	1	12.50	10.00
45.02	1	1.25	1	12.50	10.00
45.373	1	1.25	1	12.50	10.00
45.945	1	1.25	1	12.50	10.00
46.155	1	1.25	1	12.50	10.00
46.229	1	1.25	1	12.50	10.00
48.943	1	1.25	1	12.50	10.00
50.047	1	1.25	1	12.50	10.00
51.144	1	1.25	1	12.50	10.00
51.976	1	1.25	1	12.50	10.00
52.523	1	1.25	1	12.50	10.00
255.723	2	2.50	2	25.00	10.00
257.546	2	2.50	2	25.00	10.00
259.296	2	2.50	2	25.00	10.00
259.587	2	2.50	2	25.00	10.00
261.083	2	2.50	2	25.00	10.00
Total	80	100.00	80 (n = 8)	1000.00	10.00

图 16.35　“income”变量的分布频率

图 16.36 是“cost”变量组内、组间以及整体的分布频率的结果。

图 16.37 是“profit”变量组内、组间以及整体的分布频率的结果。

. xttab cost

cost	Overall Freq.	Overall Percent	Between Freq.	Between Percent	Within Percent
4.7	2	2.50	2	25.00	10.00
4.79	2	2.50	2	25.00	10.00
4.8	2	2.50	2	25.00	10.00
4.94	2	2.50	2	25.00	10.00
5.09	2	2.50	2	25.00	10.00
5.22	2	2.50	2	25.00	10.00
5.5	2	2.50	2	25.00	10.00
5.8	2	2.50	2	25.00	10.00
6	2	2.50	2	25.00	10.00
6.04	2	2.50	2	25.00	10.00
6.48	2	2.50	2	25.00	10.00
6.67	3	3.75	3	37.50	10.00
6.81	2	2.50	2	25.00	10.00
6.87	1	1.25	1	12.50	10.00
6.89	2	2.50	2	25.00	10.00
7.12	2	2.50	2	25.00	10.00
7.18	2	2.50	2	25.00	10.00
7.26	2	2.50	2	25.00	10.00
7.62	1	1.25	1	12.50	10.00
7.93	2	2.50	2	25.00	10.00
8.15	1	1.25	1	12.50	10.00
8.25	2	2.50	2	25.00	10.00
8.27	2	2.50	2	25.00	10.00
8.41	2	2.50	2	25.00	10.00
8.51	1	1.25	1	12.50	10.00
8.64	1	1.25	1	12.50	10.00
8.81	2	2.50	2	25.00	10.00
9.06	1	1.25	1	12.50	10.00
9.18	2	2.50	2	25.00	10.00
9.5	2	2.50	2	25.00	10.00
9.69	1	1.25	1	12.50	10.00
9.73	1	1.25	1	12.50	10.00
9.99	1	1.25	1	12.50	10.00
10.85	1	1.25	1	12.50	10.00
10.95	1	1.25	1	12.50	10.00
11.02	2	2.50	2	25.00	10.00
11.53	1	1.25	1	12.50	10.00
12.05	1	1.25	1	12.50	10.00
12.49	1	1.25	1	12.50	10.00
12.53	1	1.25	1	12.50	10.00
13.62	1	1.25	1	12.50	10.00
14.15	1	1.25	1	12.50	10.00
14.67	1	1.25	1	12.50	10.00
17.85	1	1.25	1	12.50	10.00
38.68	2	2.50	2	25.00	10.00
39.35	2	2.50	2	25.00	10.00
44.81	2	2.50	2	25.00	10.00
52.23	2	2.50	2	25.00	10.00
58.56	2	2.50	2	25.00	10.00
Total	80	100.00	80 (n = 8)	1000.00	10.00

图 16.36　“cost”变量的分布频率

. xttab profit

profit	Overall Freq.	Overall Percent	Between Freq.	Between Percent	Within Percent
1.026	2	2.50	2	25.00	10.00
1.195	2	2.50	2	25.00	10.00
1.211	2	2.50	2	25.00	10.00
1.495	2	2.50	2	25.00	10.00
1.656	2	2.50	2	25.00	10.00
1.665	2	2.50	2	25.00	10.00
1.779	2	2.50	2	25.00	10.00
1.811	2	2.50	2	25.00	10.00
1.903	2	2.50	2	25.00	10.00
1.923	2	2.50	2	25.00	10.00
1.94	2	2.50	2	25.00	10.00
1.953	2	2.50	2	25.00	10.00
1.992	2	2.50	2	25.00	10.00
2.023	1	1.25	1	12.50	10.00
2.063	2	2.50	2	25.00	10.00
2.226	1	1.25	1	12.50	10.00
2.322	1	1.25	1	12.50	10.00
2.346	2	2.50	2	25.00	10.00
2.531	2	2.50	2	25.00	10.00
2.575	1	1.25	1	12.50	10.00
2.579	1	1.25	1	12.50	10.00
2.582	1	1.25	1	12.50	10.00
2.61	2	2.50	2	25.00	10.00
2.634	2	2.50	2	25.00	10.00
2.638	1	1.25	1	12.50	10.00
2.685	1	1.25	1	12.50	10.00
2.758	1	1.25	1	12.50	10.00
2.933	1	1.25	1	12.50	10.00
2.936	1	1.25	1	12.50	10.00
2.996	1	1.25	1	12.50	10.00
3.042	2	2.50	2	25.00	10.00
3.108	2	2.50	2	25.00	10.00
3.148	1	1.25	1	12.50	10.00
3.247	1	1.25	1	12.50	10.00
3.275	2	2.50	2	25.00	10.00
3.406	2	2.50	2	25.00	10.00
3.5	2	2.50	2	25.00	10.00
3.533	1	1.25	1	12.50	10.00
3.553	1	1.25	1	12.50	10.00
3.695	2	2.50	2	25.00	10.00
3.738	1	1.25	1	12.50	10.00
3.835	2	2.50	2	25.00	10.00
3.884	1	1.25	1	12.50	10.00
3.9	1	1.25	1	12.50	10.00
3.975	1	1.25	1	12.50	10.00
19.536	2	2.50	2	25.00	10.00
20.201	2	2.50	2	25.00	10.00
21.039	2	2.50	2	25.00	10.00
21.268	2	2.50	2	25.00	10.00
26.211	2	2.50	2	25.00	10.00
Total	80	100.00	80	1000.00	10.00
			(n = 8)		

图 16.37　“profit”变量的分布频率

图 16.38 是对每个个体显示“income”变量的时间序列图的结果。

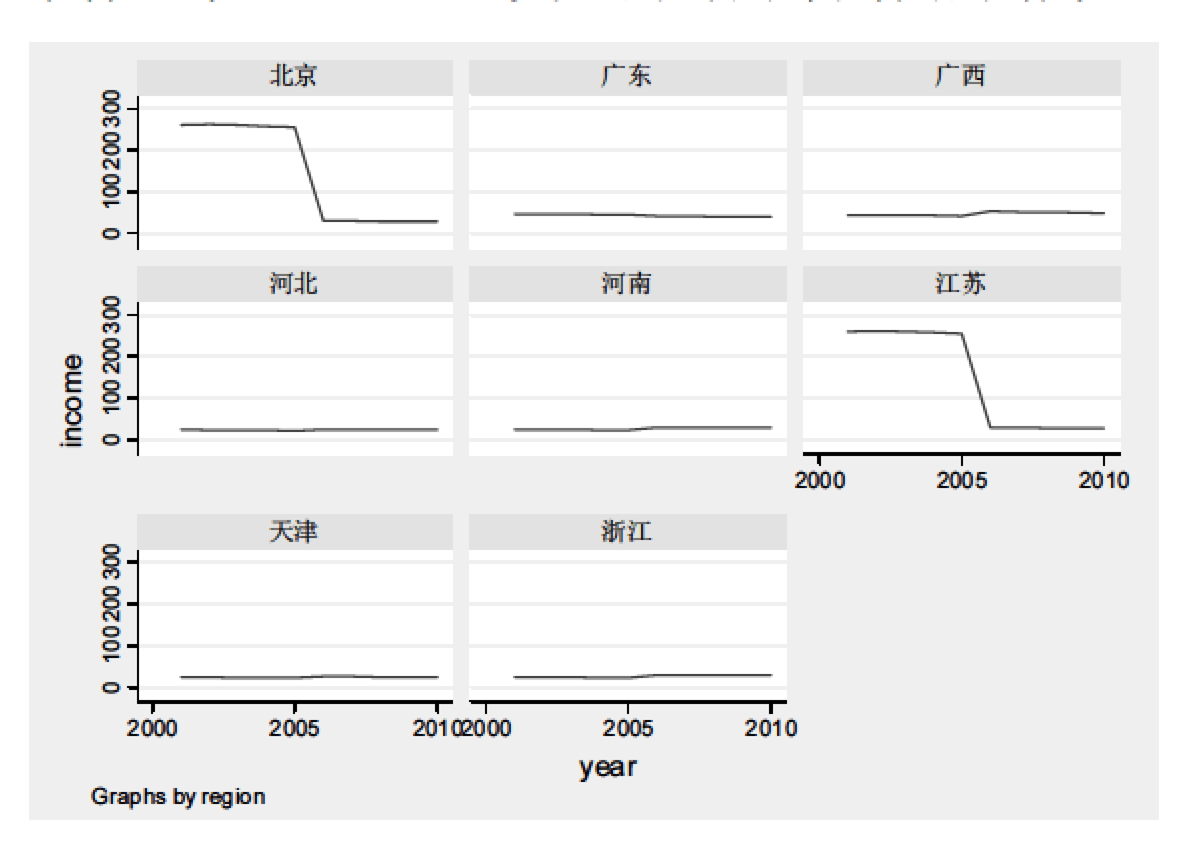


图 16.38　显示“income”变量的时间序列图

从图 16.38 可以看出，不同地区的保费收入的时间趋势是不一致的，有的地区变化一直非常平稳，有的地区先平稳再下降后平稳。

图 16.39 是对每个个体显示“cost”变量的时间序列图的结果。

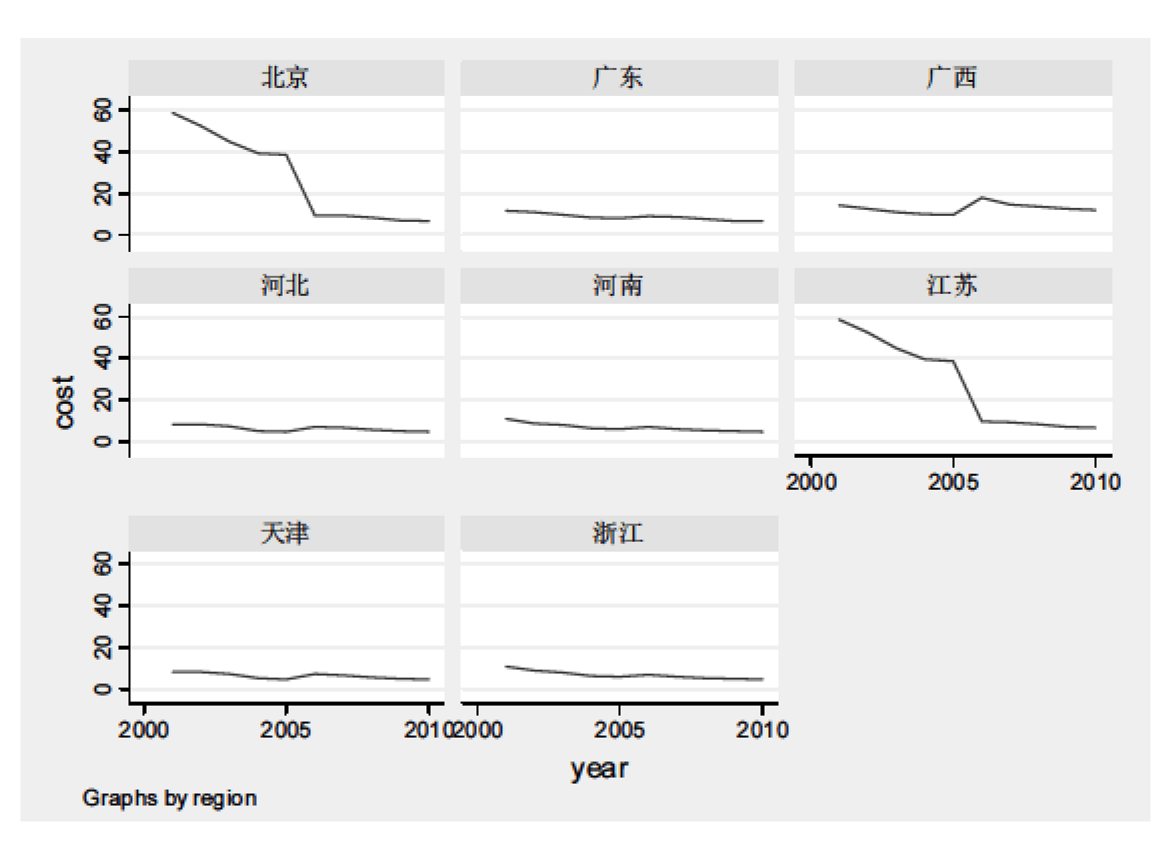


图 16.39　显示“cost”变量的时间序列图

从图 16.39 中可以看出，不同地区的赔偿支出的时间趋势是不一致的，有的地区变化一直非常平稳，有的地区先平稳再下降后平稳。

图 16.40 是对每个个体显示“profit”变量的时间序列图的结果。

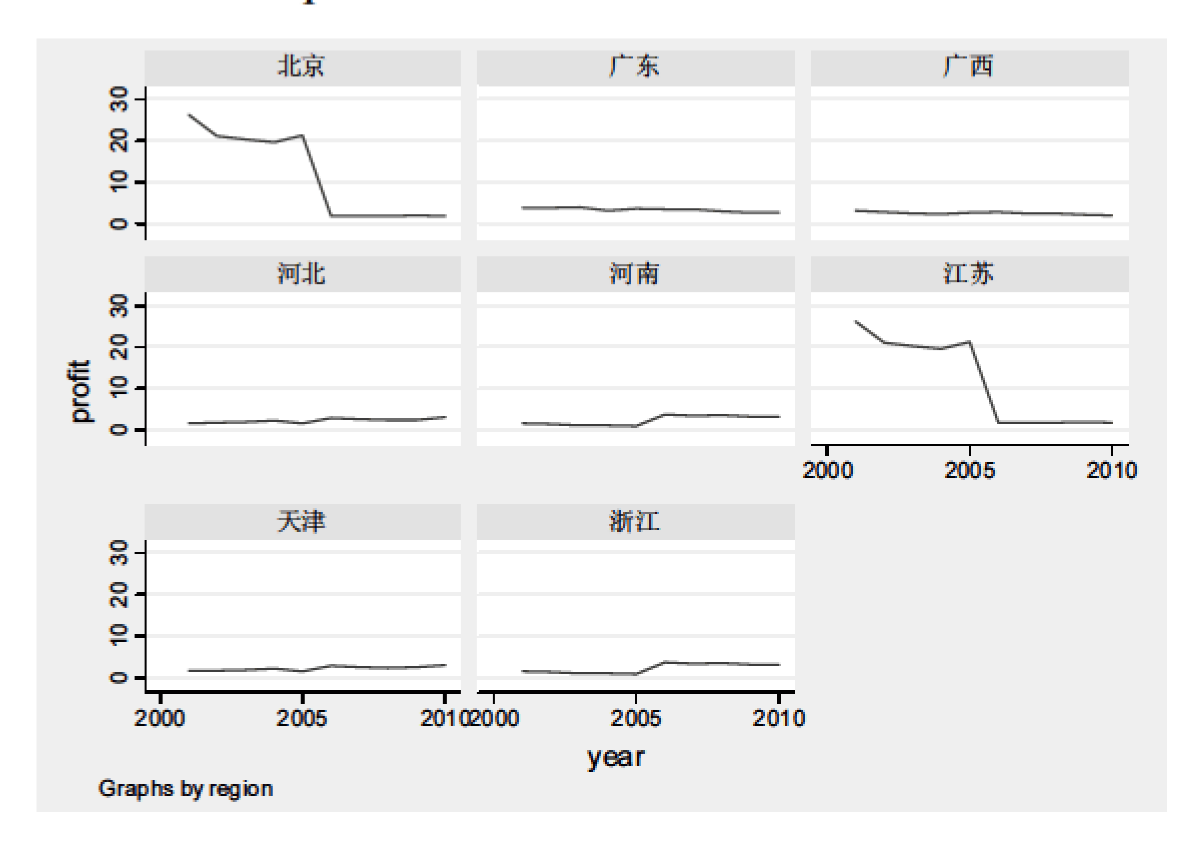
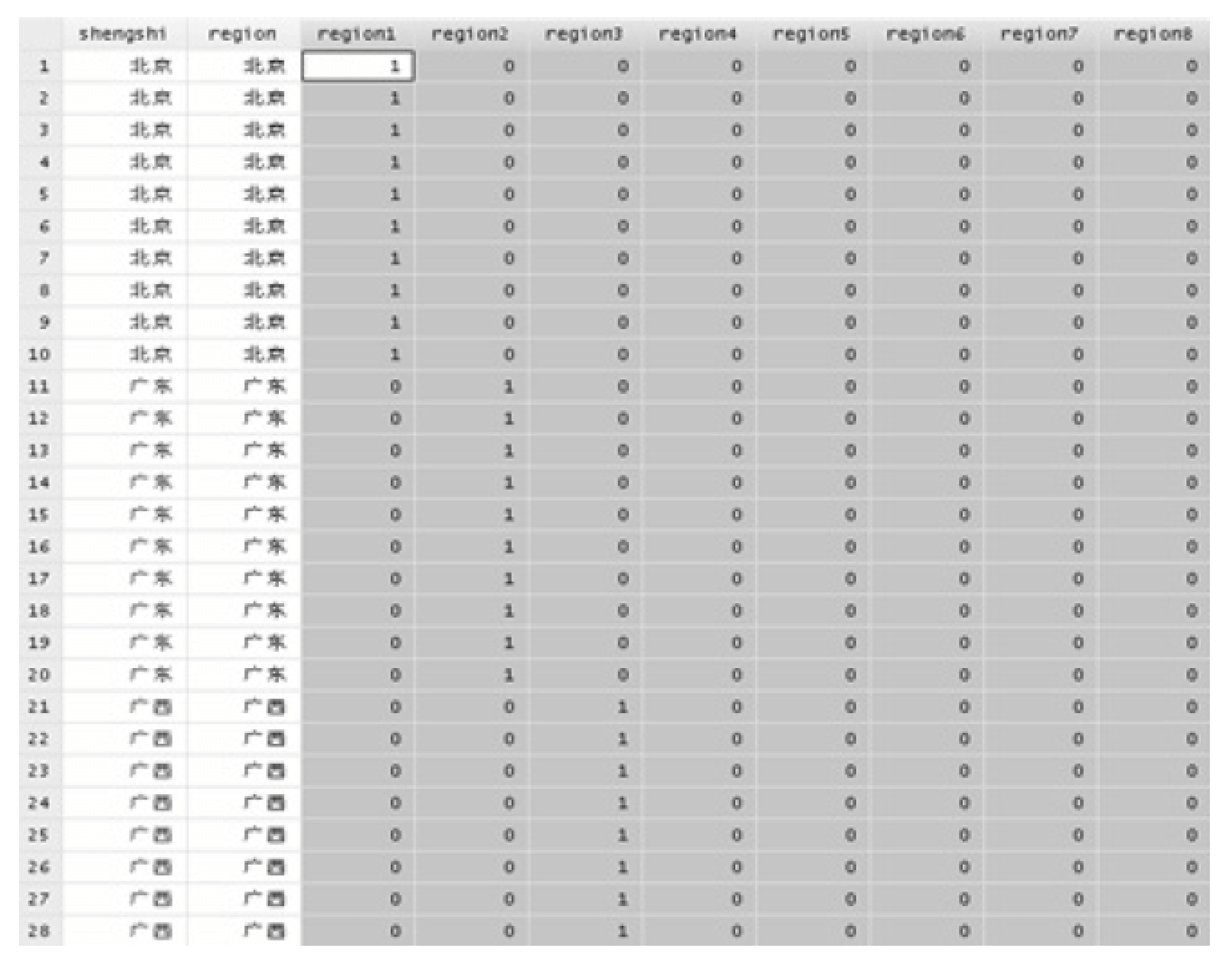


图 16.40　显示“profit”变量的时间序列图

从图 16.40 可以看出，不同地区的创造利润的时间趋势是不一致的，有的地区变化一直非常平稳，有的地区先平稳再下降后平稳。

图 16.41 是创建省市变量的多个虚拟变量的结果。选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 16.41 所示的变量 region1~region8 的相关数据。

图 16.42 是以 profit 为因变量，以 income、cost 以及生成的各个地区虚拟变量为自变量，并使用以“region”为聚类变量的聚类稳健标准差，进行最小二乘回归分析的结果。



	shengshi	region	region1	region2	region3	region4	region5	region6	region7	region8
1	北京	北京	1	0	0	0	0	0	0	0
2	北京	北京	1	0	0	0	0	0	0	0
3	北京	北京	1	0	0	0	0	0	0	0
4	北京	北京	1	0	0	0	0	0	0	0
5	北京	北京	1	0	0	0	0	0	0	0
6	北京	北京	1	0	0	0	0	0	0	0
7	北京	北京	1	0	0	0	0	0	0	0
8	北京	北京	1	0	0	0	0	0	0	0
9	北京	北京	1	0	0	0	0	0	0	0
10	北京	北京	1	0	0	0	0	0	0	0
11	广东	广东	0	1	0	0	0	0	0	0
12	广东	广东	0	1	0	0	0	0	0	0
13	广东	广东	0	1	0	0	0	0	0	0
14	广东	广东	0	1	0	0	0	0	0	0
15	广东	广东	0	1	0	0	0	0	0	0
16	广东	广东	0	1	0	0	0	0	0	0
17	广东	广东	0	1	0	0	0	0	0	0
18	广东	广东	0	1	0	0	0	0	0	0
19	广东	广东	0	1	0	0	0	0	0	0
20	广东	广东	0	1	0	0	0	0	0	0
21	广西	广西	0	0	1	0	0	0	0	0
22	广西	广西	0	0	1	0	0	0	0	0
23	广西	广西	0	0	1	0	0	0	0	0
24	广西	广西	0	0	1	0	0	0	0	0
25	广西	广西	0	0	1	0	0	0	0	0
26	广西	广西	0	0	1	0	0	0	0	0
27	广西	广西	0	0	1	0	0	0	0	0
28	广西	广西	0	0	1	0	0	0	0	0

图 16.41　创建省市变量的多个虚拟变量

```
. reg profit income cost region2-region8 year,vce(cluster region)

Linear regression                                      Number of obs =      80
                                                       F(  2,      7) =       .
                                                       Prob > F      =       .
                                                       R-squared     =  0.9845
                                                       Root MSE      =  .86123

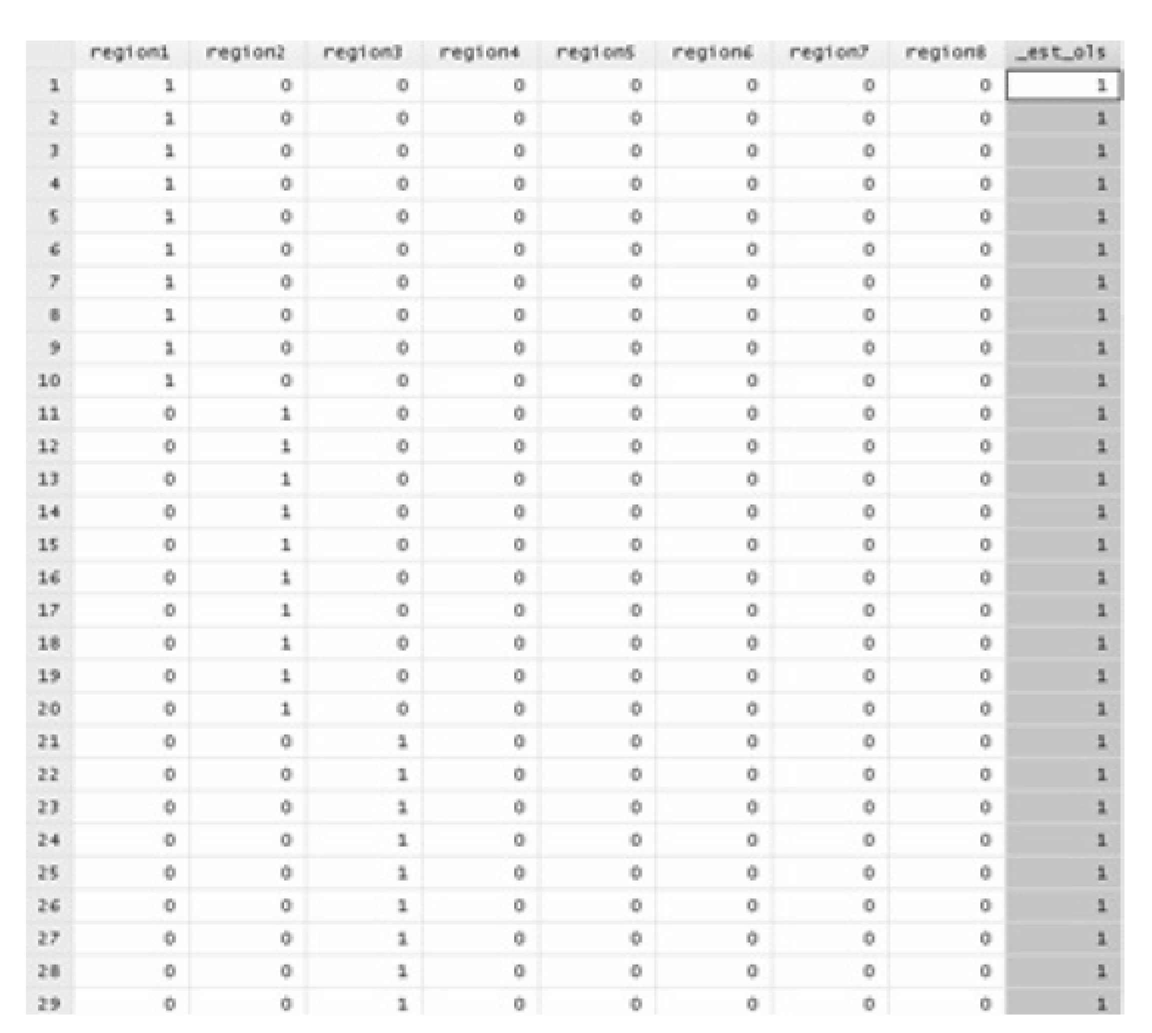
                               (Std. Err. adjusted for 8 clusters in region)
------------------------------------------------------------------------------
             |               Robust
      profit |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   .0533635   .0096339     5.54   0.001      .030583     .076144
        cost |   .2152267   .0666928     3.23   0.015     .0575234      .37293
     region2 |   1.025832   .3450825     2.97   0.021     .2098411    1.841822
     region3 |  -.8861502   .1849455    -4.79   0.002    -1.323477   -.4488235
     region4 |    1.45481   .3547113     4.10   0.005     .6160509    2.293569
     region5 |   1.280719   .3443042     3.72   0.007     .4665685    2.094869
     region6 |  -3.88e-15   6.44e-14    -0.06   0.954    -1.56e-13    1.48e-13
     region7 |    1.45481   .3547113     4.10   0.005     .6160509    2.293569
     region8 |   1.280719   .3443042     3.72   0.007     .4665685    2.094869
        year |   .1668369   .1098037     1.52   0.172    -.0928075    .4264813
       _cons |  -336.3782   220.7297    -1.52   0.171     -858.321    185.5646
------------------------------------------------------------------------------
```

图 16.42 最小二乘回归分析

从图 16.42 所示的分析结果中可以看出共有 80 个样本参与了分析。模型的可决系数（R-squared）为 0.9845，说明模型的解释能力是非常好的。

从上面的分析可以看出最小二乘线性模型的整体显著性、系数显著性以及模型的整体解释能力都很不错。得到的结论是该保险公司的创造利润情况与保费收入和赔偿支出等都是显著呈正向变化的。

图 16.43 存储的是普通最小二乘回归分析估计结果。选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 16.43 所示的变量_est_ols 的相关数据。



	region1	region2	region3	region4	region5	region6	region7	region8	_est_ols
1	1	0	0	0	0	0	0	0	1
2	1	0	0	0	0	0	0	0	1
3	1	0	0	0	0	0	0	0	1
4	1	0	0	0	0	0	0	0	1
5	1	0	0	0	0	0	0	0	1
6	1	0	0	0	0	0	0	0	1
7	1	0	0	0	0	0	0	0	1
8	1	0	0	0	0	0	0	0	1
9	1	0	0	0	0	0	0	0	1
10	1	0	0	0	0	0	0	0	1
11	0	1	0	0	0	0	0	0	1
12	0	1	0	0	0	0	0	0	1
13	0	1	0	0	0	0	0	0	1
14	0	1	0	0	0	0	0	0	1
15	0	1	0	0	0	0	0	0	1
16	0	1	0	0	0	0	0	0	1
17	0	1	0	0	0	0	0	0	1
18	0	1	0	0	0	0	0	0	1
19	0	1	0	0	0	0	0	0	1
20	0	1	0	0	0	0	0	0	1
21	0	0	1	0	0	0	0	0	1
22	0	0	1	0	0	0	0	0	1
23	0	0	1	0	0	0	0	0	1
24	0	0	1	0	0	0	0	0	1
25	0	0	1	0	0	0	0	0	1
26	0	0	1	0	0	0	0	0	1
27	0	0	1	0	0	0	0	0	1
28	0	0	1	0	0	0	0	0	1
29	0	0	1	0	0	0	0	0	1

图 16.43 普通最小二乘回归分析

图 16.44 是在仅考虑存在组内自相关，并且各组的自回归系数相同的情形下，以 profit 为因变量，以 income、cost 以及生成的各个地区虚拟变量为自变量，进行可行广义最小二乘回归分析的结果。

```
. xtpcse profit income cost region2-region8 year,corr(ar1)

Prais-Winsten regression, correlated panels corrected standard errors (PCSEs)

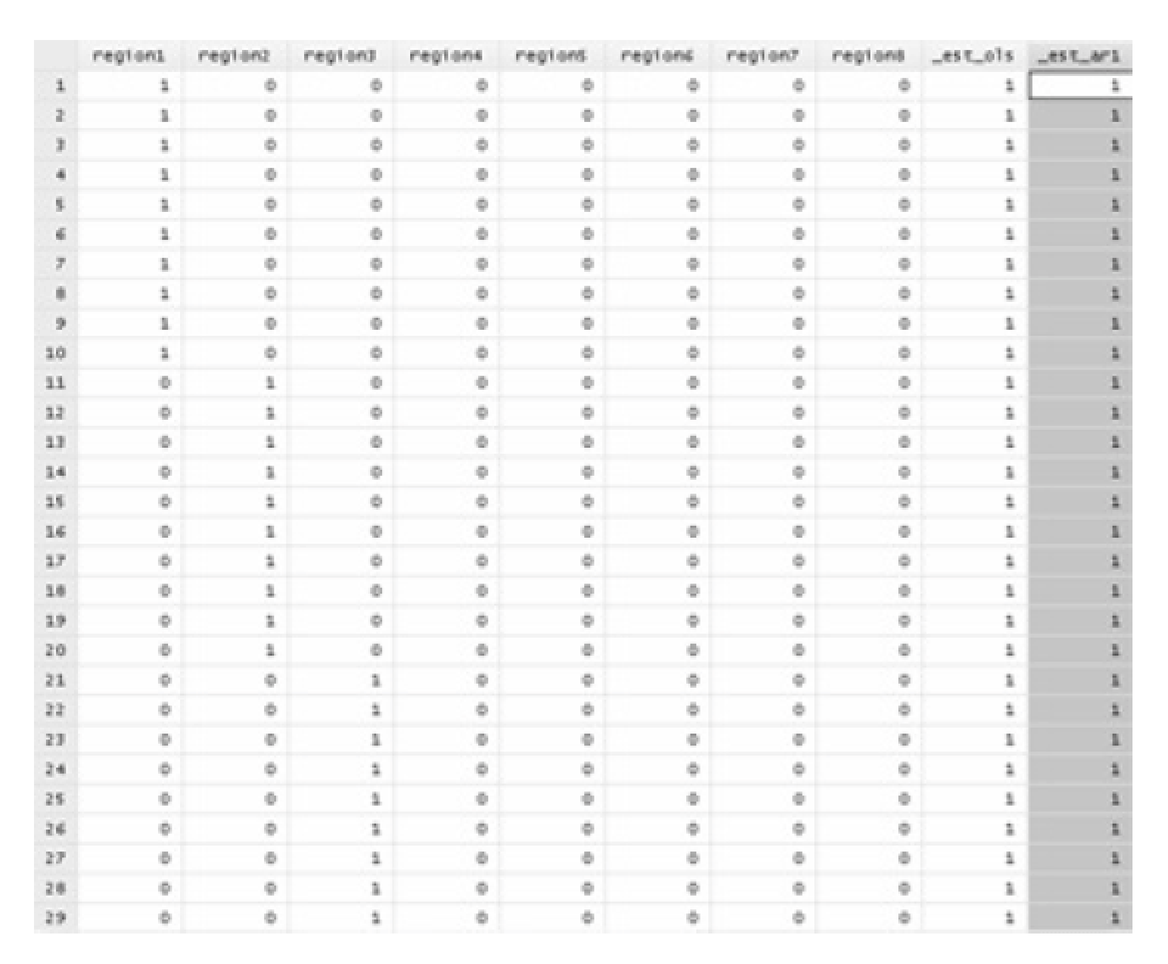
Group variable:   region                          Number of obs     =        80
Time variable:    year                            Number of groups  =         8
Panels:           correlated (balanced)           Obs per group: min =      10
Autocorrelation:  common AR(1)                                   avg =      10
                                                                 max =      10
Estimated covariances      =        36            R-squared         =    0.9794
Estimated autocorrelations =         1            Wald chi2(8)      =   1031.38
Estimated coefficients     =        11            Prob > chi2       =    0.0000

------------------------------------------------------------------------------
             |           Panel-corrected
      profit |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   .0513848   .0114491     4.49   0.000     .0289448    .0738247
        cost |   .2369246   .0685292     3.46   0.001     .1026099    .3712394
     region2 |   1.148906   .6534121     1.76   0.079    -.1317581     2.42957
     region3 |  -.8322166    .695395    -1.20   0.231    -2.195166    .5307325
     region4 |   1.610996   .6838901     2.36   0.018     .2705958    2.951396
     region5 |   1.413287   .7366856     1.92   0.055    -.0305905    2.857164
     region6 |  -1.89e-12   3.36e-08    -0.00   1.000    -6.59e-08    6.58e-08
     region7 |   1.610996   .6838901     2.36   0.018     .2705958    2.951396
     region8 |   1.413287   .7366856     1.92   0.055    -.0305905    2.857164
        year |   .1793389   .0370433     4.84   0.000     .1067353    .2519424
       _cons |  -361.6927   74.62795    -4.85   0.000    -507.9608   -215.4246
-------------+----------------------------------------------------------------
         rho |    .265627
------------------------------------------------------------------------------
```

图 16.44　进行可行广义最小二乘回归分析

从图 16.44 可以看出，在仅考虑存在组内自相关，并且各组的自回归系数相同的情形下，进行可行广义最小二乘回归分析的结果与普通最小二乘回归分析的结果是有一些区别的。

图 16.45 存储的是上步可行广义最小二乘回归分析的估计结果。选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 16.45 所示的变量_est_ar1 的相关数据。



	region1	region2	region3	region4	region5	region6	region7	region8	_est_ols	_est_ar1
1	1	0	0	0	0	0	0	0	1	1
2	1	0	0	0	0	0	0	0	1	1
3	1	0	0	0	0	0	0	0	1	1
4	1	0	0	0	0	0	0	0	1	1
5	1	0	0	0	0	0	0	0	1	1
6	1	0	0	0	0	0	0	0	1	1
7	1	0	0	0	0	0	0	0	1	1
8	1	0	0	0	0	0	0	0	1	1
9	1	0	0	0	0	0	0	0	1	1
10	1	0	0	0	0	0	0	0	1	1
11	0	1	0	0	0	0	0	0	1	1
12	0	1	0	0	0	0	0	0	1	1
13	0	1	0	0	0	0	0	0	1	1
14	0	1	0	0	0	0	0	0	1	1
15	0	1	0	0	0	0	0	0	1	1
16	0	1	0	0	0	0	0	0	1	1
17	0	1	0	0	0	0	0	0	1	1
18	0	1	0	0	0	0	0	0	1	1
19	0	1	0	0	0	0	0	0	1	1
20	0	1	0	0	0	0	0	0	1	1
21	0	0	1	0	0	0	0	0	1	1
22	0	0	1	0	0	0	0	0	1	1
23	0	0	1	0	0	0	0	0	1	1
24	0	0	1	0	0	0	0	0	1	1
25	0	0	1	0	0	0	0	0	1	1
26	0	0	1	0	0	0	0	0	1	1
27	0	0	1	0	0	0	0	0	1	1
28	0	0	1	0	0	0	0	0	1	1
29	0	0	1	0	0	0	0	0	1	1

图 16.45　查看数据

图 16.46 是在仅考虑存在组内自相关，并且各组的自回归系数不相同的情形下，以 profit 为因变量，以 income、cost 以及生成的各个地区虚拟变量为自变量，进行可行广义最小二乘回归分析的结果。

```
. xtpcse profit income cost region2-region8 year,corr(psar1)

Prais-Winsten regression, correlated panels corrected standard errors (PCSEs)

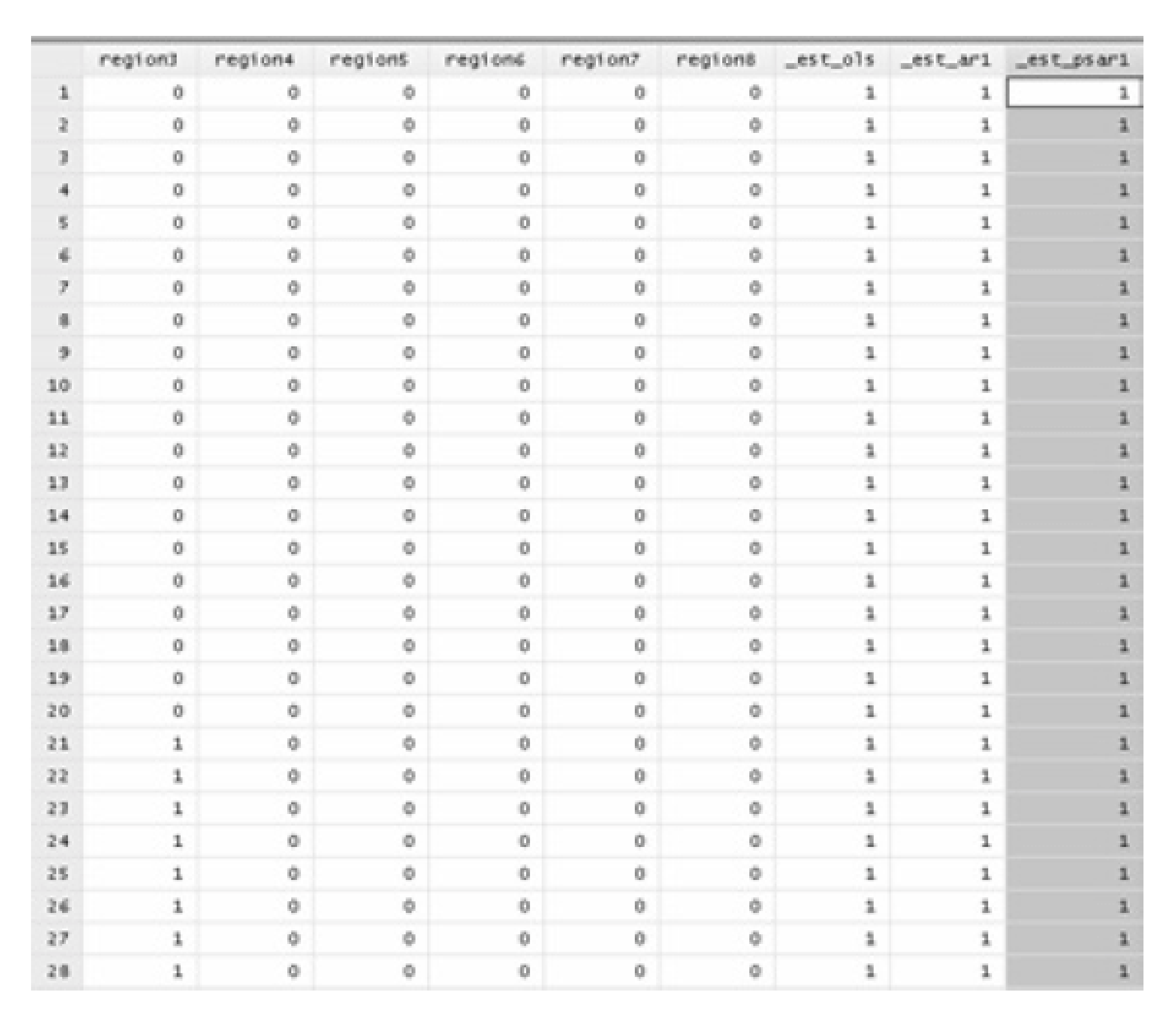
Group variable:   region                     Number of obs      =        80
Time variable:    year                       Number of groups   =         8
Panels:           correlated (balanced)      Obs per group: min =        10
Autocorrelation:  panel-specific AR(1)                      avg =        10
                                                            max =        10
Estimated covariances      =        36       R-squared          =    0.9925
Estimated autocorrelations =         8       Wald chi2(8)       =   2660.97
Estimated coefficients     =        11       Prob > chi2        =    0.0000

------------------------------------------------------------------------------
             |           Panel-corrected
      profit |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   .0499286   .0088864     5.62   0.000     .0325115    .0673457
        cost |   .2353169    .053092     4.43   0.000     .1312585    .3393753
     region2 |   .9777836   .5988821     1.63   0.103    -.1960038    2.151571
     region3 |  -.9068021   .7989255    -1.14   0.256    -2.472667    .6590631
     region4 |   1.504788   .4177599     3.60   0.000      .685994    2.323583
     region5 |   1.276868   .6133663     2.08   0.037     .0746926    2.479044
     region6 |   6.31e-13   2.32e-08     0.00   1.000    -4.56e-08    4.56e-08
     region7 |   1.504788   .4177599     3.60   0.000      .685994    2.323583
     region8 |   1.276868   .6133663     2.08   0.037     .0746926    2.479044
        year |   .1973701   .0359409     5.49   0.000     .1269273    .2678129
       _cons |  -397.7056   72.26995    -5.50   0.000    -539.3521   -256.0591
-------------+----------------------------------------------------------------
        rhos = -.1981808   .8593703   .7428073 -.1559056   .6155057  ...  -.1981808
------------------------------------------------------------------------------
```

图 16.46　自回归系数不相同

从图 16.46 可以看出，在仅考虑存在组内自相关，并且各组的自回归系数不相同的情形下，进行可行广义最小二乘回归分析的结果与前面各种回归分析的结果是有一些区别的。

图 16.47 存储的是上步可行广义最小二乘回归分析的估计结果。选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 16.47 所示的变量_est_psar1 的相关数据。



	region3	region4	region5	region6	region7	region8	_est_ols	_est_ar1	_est_psar1
1	0	0	0	0	0	0	1	1	1
2	0	0	0	0	0	0	1	1	1
3	0	0	0	0	0	0	1	1	1
4	0	0	0	0	0	0	1	1	1
5	0	0	0	0	0	0	1	1	1
6	0	0	0	0	0	0	1	1	1
7	0	0	0	0	0	0	1	1	1
8	0	0	0	0	0	0	1	1	1
9	0	0	0	0	0	0	1	1	1
10	0	0	0	0	0	0	1	1	1
11	0	0	0	0	0	0	1	1	1
12	0	0	0	0	0	0	1	1	1
13	0	0	0	0	0	0	1	1	1
14	0	0	0	0	0	0	1	1	1
15	0	0	0	0	0	0	1	1	1
16	0	0	0	0	0	0	1	1	1
17	0	0	0	0	0	0	1	1	1
18	0	0	0	0	0	0	1	1	1
19	0	0	0	0	0	0	1	1	1
20	0	0	0	0	0	0	1	1	1
21	1	0	0	0	0	0	1	1	1
22	1	0	0	0	0	0	1	1	1
23	1	0	0	0	0	0	1	1	1
24	1	0	0	0	0	0	1	1	1
25	1	0	0	0	0	0	1	1	1
26	1	0	0	0	0	0	1	1	1
27	1	0	0	0	0	0	1	1	1
28	1	0	0	0	0	0	1	1	1

图 16.47　查看数据

图 16.48 是在不考虑存在自相关，仅考虑不同个体扰动项存在异方差的情形下，以 profit 为因变量，以 income、cost 以及生成的各个地区虚拟变量为自变量，进行可行广义最小二乘回归分析的结果。

```
. xtpcse profit income cost region2-region8 year,hetonly

Linear regression, heteroskedastic panels corrected standard errors

Group variable:   region                          Number of obs      =        80
Time variable:    year                            Number of groups   =         8
Panels:           heteroskedastic (balanced)      Obs per group: min =        10
Autocorrelation:  no autocorrelation                             avg =        10
                                                                 max =        10

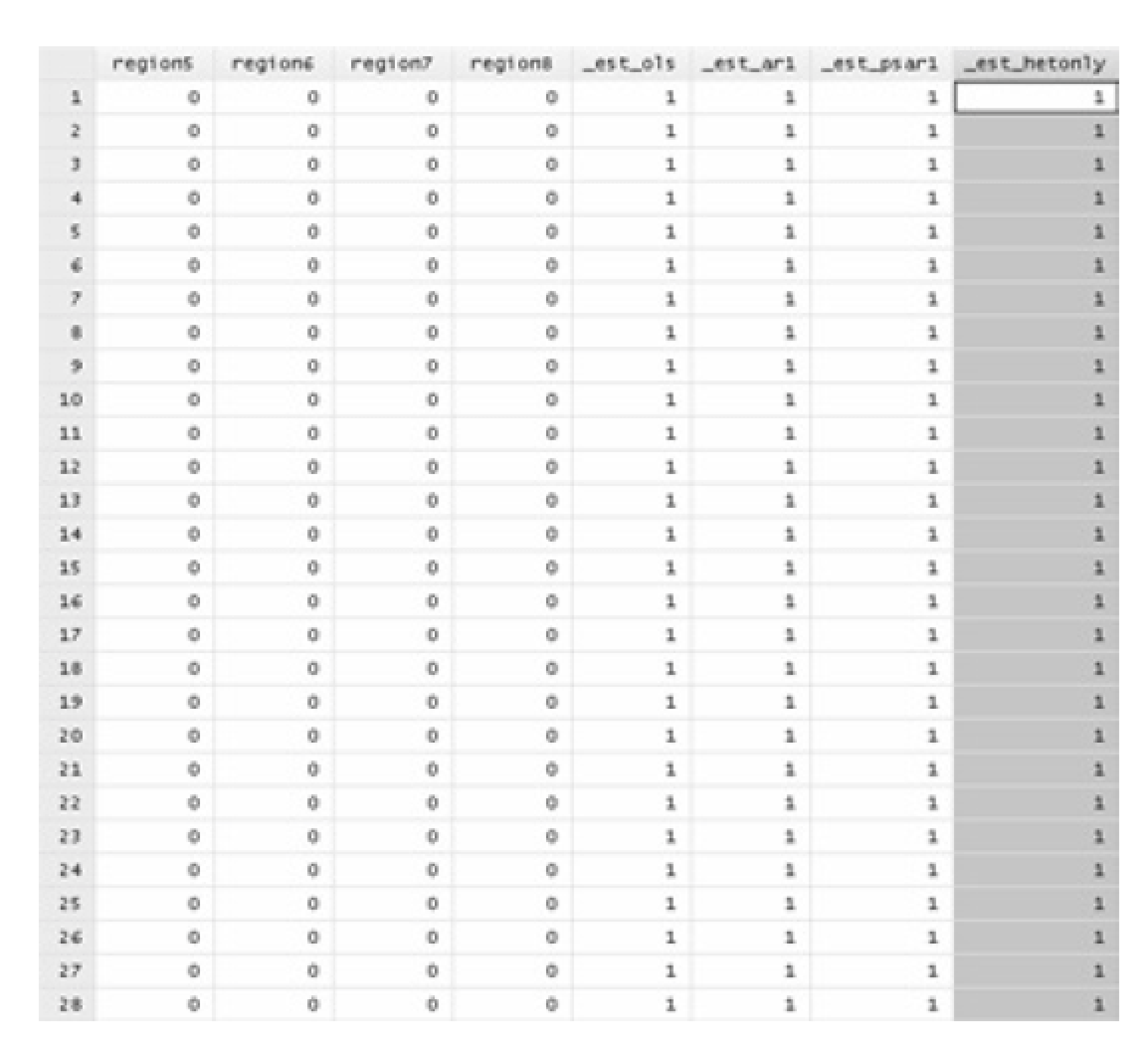
Estimated covariances      =         8            R-squared          =    0.9845
Estimated autocorrelations =         0            Wald chi2(10)      =   3241.67
Estimated coefficients     =        11            Prob > chi2        =    0.0000

------------------------------------------------------------------------------
             |           Het-corrected
      profit |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   .0533635   .0073228     7.29   0.000     .0390111    .0677159
        cost |   .2152267   .0444006     4.85   0.000     .1282031    .3022503
     region2 |   1.025832   .4483788     2.29   0.022     .1470253    1.904638
     region3 |  -.8861502   .5174744    -1.71   0.087    -1.900381    .1280809
     region4 |    1.45481   .4465286     3.26   0.001     .5796298     2.32999
     region5 |   1.280719   .5055611     2.53   0.011      .289837      2.2716
     region6 |  -3.88e-15   .4762843    -0.00   1.000    -.9335001    .9335001
     region7 |    1.45481   .4465286     3.26   0.001     .5796298     2.32999
     region8 |   1.280719   .5055611     2.53   0.011      .289837      2.2716
        year |   .1668369    .038223     4.36   0.000     .0919212    .2417526
       _cons |  -336.3782   76.85813    -4.38   0.000    -487.0174   -185.7391
------------------------------------------------------------------------------
```

图 16.48　仅考虑不同个体扰动项存在异方差

从图 16.48 可以看出，在不考虑存在自相关，仅考虑不同个体扰动项存在异方差的情形下，进行可行广义最小二乘回归分析的结果与前面各种回归分析的结果是有一些区别的。

图 16.49 存储的是上步可行广义最小二乘回归分析的估计结果。选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 16.49 所示的变量_est_ hetonly 的相关数据。



	region5	region6	region7	region8	_est_ols	_est_ar1	_est_psar1	_est_hetonly
1	0	0	0	0	1	1	1	1
2	0	0	0	0	1	1	1	1
3	0	0	0	0	1	1	1	1
4	0	0	0	0	1	1	1	1
5	0	0	0	0	1	1	1	1
6	0	0	0	0	1	1	1	1
7	0	0	0	0	1	1	1	1
8	0	0	0	0	1	1	1	1
9	0	0	0	0	1	1	1	1
10	0	0	0	0	1	1	1	1
11	0	0	0	0	1	1	1	1
12	0	0	0	0	1	1	1	1
13	0	0	0	0	1	1	1	1
14	0	0	0	0	1	1	1	1
15	0	0	0	0	1	1	1	1
16	0	0	0	0	1	1	1	1
17	0	0	0	0	1	1	1	1
18	0	0	0	0	1	1	1	1
19	0	0	0	0	1	1	1	1
20	0	0	0	0	1	1	1	1
21	0	0	0	0	1	1	1	1
22	0	0	0	0	1	1	1	1
23	0	0	0	0	1	1	1	1
24	0	0	0	0	1	1	1	1
25	0	0	0	0	1	1	1	1
26	0	0	0	0	1	1	1	1
27	0	0	0	0	1	1	1	1
28	0	0	0	0	1	1	1	1

图 16.49　查看数据

图 16.50 是展示将以上各种方法的系数估计值及标准差列表放到一起进行比较的结果。

```
. estimates table ols ar1 psar1 hetonly,b se

----------------------------------------------------------------
    Variable |    ols          ar1          psar1        hetonly
-------------+--------------------------------------------------
      income |  .05336351    .05138476    .04992861    .05336351
             |  .00963388    .01144915    .00888643     .0073228
        cost |   .2152267    .23692465    .23531693     .2152267
             |  .06669277    .06852918    .05309199    .04440063
     region2 |  1.0258316     1.148906    .97778357    1.0258316
             |  .34508253    .65341206    .59888209    .44837881
     region3 | -.88615016   -.83221655   -.90680209   -.88615016
             |   .1849455    .69539496    .79892547    .51747435
     region4 |  1.4548098    1.6109958    1.5047883    1.4548098
             |  .35471127    .68389013    .41775991    .44652861
     region5 |  1.2807187    1.4132868    1.2768684    1.2807187
             |  .34430425     .7366856    .61336628    .50556113
     region6 | -3.883e-15   -1.886e-12    6.306e-13   -3.883e-15
             |  6.440e-14    3.360e-08    2.325e-08    .47628431
     region7 |  1.4548098    1.6109958    1.5047883    1.4548098
             |  .35471127    .68389013    .41775991    .44652861
     region8 |  1.2807187    1.4132868    1.2768684    1.2807187
             |  .34430425     .7366856    .61336628    .50556113
        year |  .16683689    .17933885    .19737013    .16683689
             |  .10980365    .03704331    .03594086    .03822298
       _cons | -336.37823   -361.69267    -397.7056   -336.37823
             |   220.7297    74.627951    72.269954    76.858126
----------------------------------------------------------------
                                                     legend: b/se
```

图 16.50　展示比较结果

从图16.50可以看出，hetonly方法的系数估计值和ols方法的系数估计值是完全一样的，但是标准差并不一样。其他各种方法之间都存在着一定的差别。

图 16.51 是在假定不同个体的扰动项相互独立且有不同的方差，并且各组的自回归系数相同的情形下，以 profit 为因变量，以 income、cost 以及生成的各个地区虚拟变量为自变量，进行可行广义最小二乘回归分析的结果。

```
. xtgls profit income cost region2-region8 year,panels(cor) cor(ar1)

Crcss-sectional time-series FGLS regression

Coefficients:  generalized least squares
Parels:        heteroskedastic with cross-sectional correlation
Correlation:   common AR(1) coefficient for all panels  (0.2656)

Estimated covariances      =        36          Number of obs      =        80
Estimated autocorrelations =         1          Number of groups   =         8
Estimated coefficients     =         8          Time periods       =        10
                                                Wald chi2(7)       =   1144.31
                                                Prob > chi2        =    0.0000

------------------------------------------------------------------------------
      profit |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |    .050533   .0059673     8.47   0.000     .0388372    .0622288
        cost |   .2372836   .0283261     8.38   0.000     .1817655    .2928017
     region2 |   1.069898   .6140896     1.74   0.081    -.1336956    2.273491
     region3 |  -.9093056   .6905757    -1.32   0.188    -2.262809    .4441979
     region4 |          0  (omitted)
     region5 |   1.321584   .7099093     1.86   0.063    -.0698127    2.712981
     region6 |          0  (omitted)
     region7 |    1.51725   .6519075     2.33   0.020     .2395351    2.794965
     region8 |          0  (omitted)
        year |   .1623514   .0183037     8.87   0.000     .1264768     .198226
       _cons |  -327.5118   36.82939    -8.89   0.000    -399.6961   -255.3275
------------------------------------------------------------------------------
```

图 16.51　各组的自回归系数相同

从图 16.51 可以看出，在假定不同个体的扰动项相互独立且有不同的方差，并且各组的自回归系数相同的情形下，进行可行广义最小二乘回归分析的结果与前面各种回归分析的结果是有一些区别的。

图 16.52 是在假定不同个体的扰动项相互独立且有不同的方差，并且各组的自回归系数不相同的情形下，以 profit 为因变量，以 income、cost 以及生成的各个地区虚拟变量为自变量，进行可行广义最小二乘回归分析的结果。

```
. xtgls profit income cost region2-region8 year,panels(cor) cor(ar1)

Cross-sectional time-series FGLS regression

Coefficients:  generalized least squares
Panels:        heteroskedastic with cross-sectional correlation
Correlation:   common AR(1) coefficient for all panels  (0.2656)

Estimated covariances      =        36          Number of obs      =        80
Estimated autocorrelations =         1          Number of groups   =         8
Estimated coefficients     =         8          Time periods       =        10
                                                Wald chi2(7)       =   1144.31
                                                Prob > chi2        =    0.0000

------------------------------------------------------------------------------
      profit |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |    .050533   .0059673     8.47   0.000     .0388372    .0622288
        cost |   .2372836   .0283261     8.38   0.000     .1817655    .2928017
     region2 |   1.069898   .6140896     1.74   0.081    -.1336956    2.273491
     region3 |  -.9093056   .6905757    -1.32   0.188    -2.262809    .4441979
     region4 |          0  (omitted)
     region5 |   1.321584   .7099093     1.86   0.063    -.0698127    2.712981
     region6 |          0  (omitted)
     region7 |    1.51725   .6519075     2.33   0.020     .2395351    2.794965
     region8 |          0  (omitted)
        year |   .1623514   .0183037     8.87   0.000     .1264768     .198226
       _cons |  -327.5118   36.82939    -8.89   0.000    -399.6961   -255.3275
------------------------------------------------------------------------------
```

图 16.52　各组的自回归系数不相同

从图 16.52 可以看出，在假定不同个体的扰动项相互独立且有不同的方差，并且各组的自回归系数不相同的情形下，进行可行广义最小二乘回归分析的结果与前面各种回归分析的结果是有一些区别的。

16.2.5　案例延伸

上述的 Stata 命令比较简洁，分析过程及结果已达到解决实际问题的目的。但是 Stata 14.0 的强大之处在于，它同样提供了更加复杂的命令格式以满足用户更加个性化的需求。

延伸：进行随机系数模型回归分析

前面我们讲述的种种面板数据回归分析方法，最多允许每个个体拥有自己的截距项，从来没有允许每个个体拥有自己的回归方程斜率，那么 Stata 能否做到变系数呢？以本节中提到的案例为例，操作命令就是：

```
xtrc profit income cost,betas
```

本命令不仅允许每个个体拥有自己的截距项，还允许每个个体拥有自己的回归方程斜率，旨在进行随机系数模型回归分析。

在命令窗口输入命令并按回车键进行确认，结果如图 16.53 所示。

```
. xtrc profit income cost,betas

Random-coefficients regression                  Number of obs      =        80
Group variable: region                          Number of groups   =         8

                                                Obs per group: min =        10
                                                               avg =      10.0
                                                               max =        10

                                                Wald chi2(2)       =     51.09
                                                Prob > chi2        =    0.0000

      profit       Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]

      income    .1931546   .0718222     2.69   0.007     .0523857    .3339235
        cost    .0588612   .0666521     0.88   0.377    -.0717746    .1894969
       _cons   -3.104323   1.557598    -1.99   0.046    -6.157159   -.0514874

Test of parameter constancy:    chi2(21) =   891.48      Prob > chi2 = 0.0000

                        Group-specific coefficients

                   Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]

Group 1

      income    .0455572   .0059147     7.70   0.000     .0339646    .0571498
        cost    .2303642   .0476863     4.83   0.000     .1369008    .3238277
       _cons   -.9026935   .5092535    -1.77   0.076    -1.900812     .095425

Group 2

      income    .0504041    .014856     3.39   0.001     .0212869    .0795213
        cost    .1922436   .0281024     6.84   0.000     .1371639    .2473234
       _cons   -.4299306    .853226    -0.50   0.614    -2.102223    1.242362

Group 3

      income   -.0890295   .0062266   -14.30   0.000    -.1012335   -.0768256
        cost    .1874995   .0262255     7.15   0.000     .1360985    .2389005
       _cons    4.387642   .4231777    10.37   0.000     3.558229    5.217055

Group 4

      income    .3827678   .0602186     6.36   0.000     .2647415    .5007941
        cost   -.1505261   .0578134    -2.60   0.009    -.2638383   -.0372139
       _cons   -6.185409   1.167776    -5.30   0.000    -8.474208   -3.896611

Group 5

      income    .3636063   .0149761    24.28   0.000     .3342537    .3929588
        cost    -.034265   .0209232    -1.64   0.101    -.0752737    .0067437
       _cons   -7.308046    .417541   -17.50   0.000    -8.126412   -6.489681

Group 6

      income    .0455572   .0059147     7.70   0.000     .0339646    .0571498
        cost    .2303642   .0476863     4.83   0.000     .1369008    .3238277
       _cons   -.9026935   .5092535    -1.77   0.076    -1.900812     .095425

Group 7

      income    .3827678   .0602186     6.36   0.000     .2647415    .5007941
        cost   -.1505261   .0578134    -2.60   0.009    -.2638383   -.0372139
       _cons   -6.185409   1.167776    -5.30   0.000    -8.474208   -3.896611

Group 8

      income    .3636063   .0149761    24.28   0.000     .3342537    .3929588
        cost    -.034265   .0209232    -1.64   0.101    -.0752737    .0067437
       _cons   -7.308046    .417541   -17.50   0.000    -8.126412   -6.489681
```

图 16.53　分析结果图

在图 16.53 中，模型中对参数一致性检验的显著性 P 值为 0.0000（Test of parameter constancy: chi2(21) = 891.48 Prob > chi2 = 0.0000），显著地拒绝了每个个体都具有相同系数的原假设，我们的变系数模型设置是非常合理的。

可以根据上面的结果写出模型整体的回归方程和每个个体的回归方程。结果的详细解读方式与普通的最小二乘回归分析类似，限于篇幅不再赘述。

16.3　本章习题

（1）X 公司是一家销售家具的连锁公司，经营范围遍布全国 20 个省市，各省市连锁店 2008－2012 年的相关销售数据包括销售收入、促销费用以及创造利润等，如表 16.3 所示。试用多种短面板数据回归分析方法深入研究销售收入和促销费用对创造利润的影响关系。

表 16.3　X 公司各省市连锁店销售收入、促销费用以及创造利润数据（2008－2012 年）

年份	销售收入/万元	促销费用/万元	创造利润/万元	地区
2008	224.373	10.778 96	10.344 32	湖北
2009	224.723 5	11.107 96	10.178 84	湖北
2010	224.728 9	11.181 64	10.322 86	湖北
2011	224.587 7	10.968 2	10.138 96	湖北
2012	224.476 1	10.837 62	10.169 69	湖北
2008	231.01	11.699 4	9.914 922	河南

（续表）

年份	销售收入/万元	促销费用/万元	创造利润/万元	地区
…	…	…	…	…
2012	223.525 1	11.008 74	9.236 008	广东
2008	226.230 7	10.915 09	10.517 32	广西
2009	226.133 4	10.807 71	10.435 88	广西
2010	226.408 4	11.140 41	10.554 51	广西
2011	226.311 4	11.002 1	10.463 1	广西
2012	226.047 5	10.776 87	10.396 66	广西

（2）Y 公司是一家商业银行，经营范围遍布全国 10 个省市，各省市连锁店 2001－2010 年的相关经营数据包括利息收入、利息支出以及创造利润等，如表 16.4 所示。试用多种长面板数据回归分析方法深入研究利息收入、利息支出对创造利润的影响关系。

表 16.4　Y 公司各省市利息收入、利息支出以及创造利润数据（2001－2010 年）

年份	利息收入/万元	利息支出/万元	创造利润/万元	省市
2001	25.308	11.02	1.656	浙江
2002	25.281	8.81	1.495	浙江
2003	24.779	7.93	1.211	浙江
2004	24.02	6.48	1.195	浙江
2005	23.154	6.04	1.026	浙江
2006	30.892	6.89	3.835	浙江
…	…	…	…	…
2005	42.122	9.69	2.638	广西
2006	52.523	17.85	2.936	广西
2007	51.976	14.67	2.582	广西
2008	51.144	13.62	2.579	广西
2009	50.047	12.53	2.226	广西
2010	48.943	14.05	2.023	广西

第 17 章　Stata 在研究城市综合经济实力中的应用

改革开放以来，随着工业化进程的加快，我国城市的数量不断增加，个体的规模不断扩大，在社会经济生活中所起的主导作用也越来越显著。当今世界已经进入了全球经济一体化的时代，城市作为国家的经济、政治、科技和教育文化发展中心已经成为经济循环的主角，而决定每个城市的地位、作用以及未来发展态势的主要因素是它们各自拥有的综合经济实力。城市综合实力是指一个城市在一定时期内经济、社会、基础设施、环境、科技、文教等各个领域所具备的现实实力和发展能力的集合。Stata 软件可以用来进行城市综合经济实力的相关分析研究，下面我们就来介绍一下 Stata 在研究城市综合经济实力中的应用。

17.1　研究背景及目的

2009 年 10 月 17 日，第六届中国城市论坛北京峰会在朝阳规划艺术馆召开。这次峰会不仅吸引了城市发展领域内几百位专家的关注和参与，更让来自全国各个城市的会议代表们受益匪浅。会议指出，“十二五”期间既是全球经济复苏的关键时期，也是我国加快城市化进程的关键时期。

以前我国采取的城市外延式扩张战略导致城市发展中出现了资源浪费、环境污染、不注重保护城市历史文脉和特点等各种各样的问题，所以“十二五”期间，城市必须从规模、质量、结构和效益等各个角度，推进实施“内涵式发展”模式。城市发展将呈现 5 个新变化：城市发展开始从外延式扩张向内涵式发展转变；城市软实力成为城市发展的核心竞争力；城乡统筹和城乡一体化成为城市发展的新格局；综合配套改革实验区的示范意义进一步凸显；城市群对城市建设与发展的作用日益增强。

在这种大背景下对我国各城市的综合经济实力进行研究，不论是对于促进我国城市本身更加又好又快地发展，还是对于充分发挥城市在社会经济生活中所起的主导作用，都有着极为重要的意义。

本章的研究目的如下：通过对描述我国各城市综合经济实力的各种指标进行分析，一方面找出用来衡量我国城市综合经济实力的各个指标之间的内在联系，另一方面找出各城市综合经济实力的差异。

17.2　研究方法

对城市综合经济实力的概念，中国城市经济发展研究中心提出：城市综合经济实力是指

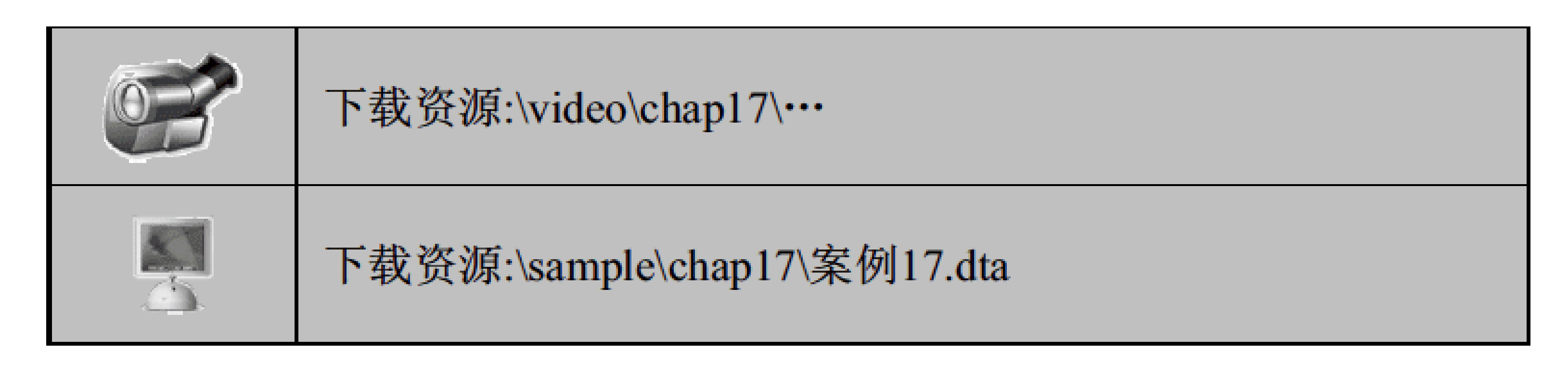
城市所拥有的全部实力、潜力及其在国内外经济社会中的地位和影响力。据此概念可以看出，评价城市综合经济实力应该包括人口、地区生产总值、拥有的交通运输以及通信能力、地方财政预算内收支、固定资产投资总额、城乡居民工资水平及储蓄水平、环境污染治理投资总额、商贸市场水平、人才状况及社会医疗保障水平等方面，所以我们采用的数据指标有：年底总人口、地区生产总值、第一产业增加值、第二产业增加值、第三产业增加值、客运量、货运量、地方财政预算内收入、地方财政预算内支出、固定资产投资总额、城乡居民储蓄年末余额、在岗职工平均工资、年末邮政局数、年末固定电话用户数、社会商品零售总额、货物进出口总额、年末实有公共汽车营运车辆数、影剧院数、普通高等学校在校学生数、医院数、执业医师、环境污染治理投资总额等 22 个指标。

本例采用的数据是《中国 2007 年省会城市和计划单列市主要经济指标统计（包括市辖县）》，数据摘编自《中国统计年鉴 2008》。

采用的数据分析方法主要有回归分析、相关分析、因子分析等。

基本思路是：首先使用回归分析、相关分析等方法研究构成城市综合经济实力的各个变量之间的关系；然后使用因子分析对构成城市综合经济实力的各个变量提取公因子；最后使用一些简单的 Stata 数据处理技巧依照提取的公因子对各城市进行分类及排序。

17.3　数据分析与报告



因为本例采用的是现有的数据，所以根据第 1 章介绍的方法直接将所用数据录入 Stata 中即可。我们共设置了 23 个变量，分别是“城市名称”“年底总人口”“地区生产总值”“第一产业增加值”“第二产业增加值”“第三产业增加值”“客运量”“货运量”“地方财政预算内收入”“地方财政预算内支出”“固定资产投资总额”“城乡居民储蓄年末余额”“在岗职工平均工资”“年末邮政局数”“年末固定电话用户数”“社会商品零售总额”“货物进出口总额”“年末实有公共汽车营运车辆数”“影剧院数”“普通高等学校在校学生数”“医院数”“执业医师”“环境污染治理投资总额”等。我们把这 23 个变量分别定义为 V1~V23。样本是中国 2007 年省会城市和计划单列市主要经济指标统计的相关数据。录入完成后数据如图 17.1 所示。

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
1	北京	1213	93533200	1012600	25094000	67426600	20040	19895	14926380	16495023
2	天津	959	50504000	1101900	28925300	20476800	7104	50462	5404390	6743262
3	石家	955	23607230	2757361	11644087	9205782	14887	12615	958720	1631692
4	太原	355	12549447	196389	6427006	5926052	3994	20126	884170	1558029
5	呼和	221	11011331	621414	4155029	6234888	4684	7976	579618	1005133
6	沈阳	710	32211508	1661506	15557413	14992589	10060	19092	2308085	3396754
7	大连	578	31306789	2492933	15355356	13458500	15446	30373	2679757	3445728
8	长春	746	20890859	2000272	10493643	8396944	7468	11134	932951	1815633
9	哈尔	987	24368044	3476764	9025670	11865610	9454	10772	1320495	2323833
10	上海	1379	1.219e+08	1018400	56785100	64085000	10370	78108	20744792	21816780
11	南京	617	32837300	864400	16072200	15900700	24594	19861	3301883	3429386
12	杭州	672	41001722	1634719	20569336	18797667	28025	22569	3916195	3357153
13	宁波	565	34350042	1512772	18990987	13846283	30693	22931	3291218	3710400
14	合肥	479	13346102	772808	6536085	6037209	8923	7186	1693175	1670789
15	福州	630	19745818	2042409	9178181	8525228	12446	12573	1465641	1430922
16	厦门	167	13878520	185113	7369519	6323888	5749	4582	1865262	1986559
17	南昌	491	13898920	867328	7542682	5488910	5531	5229	872199	1168596
18	济南	605	25628135	1502995	11580397	12544743	15388	17329	1570192	1799787
19	青岛	758	37865156	2035865	19535500	16293791	21349	40758	2928037	3211777
20	郑州	707	24867470	793847	13145716	10927907	16018	11482	2195183	2406758
21	武汉	828	31419048	1291547	14400000	15727501	17338	22552	2216755	3072345
22	长沙	637	21902548	1387971	9848274	10666303	11919	16184	1745761	2181733
23	广州	773	71091814	1498737	28067628	41525449	51180	45142	5237862	6236917
24	深圳	212	68015706	69412	34047608	33898686	15030	13754	6580554	7279563
25	南宁	684	10690099	1579371	3722713	5388015	10531	9237	701510	1180007
26	海口	153	3936858	268138	1111209	2557511	17441	6609	448495	432058
27	重庆	3235	41225100	4823900	18921000	17480200	77727	49970	4427000	7683886
28	成都	1112	33241677	2350971	15040218	15850488	43317	30026	2863772	3560453

图 17.1 案例 17 数据

17.4 描述性分析

本案例的数据变量除了城市这一字符串变量外都是定距变量，通过进行定距变量的基本描述性统计，我们可以得到数据的概要统计指标，包括平均值、最大值、最小值、标准差、百分位数、中位数、偏度系数和峰度系数等。通过获得这些指标，可以从整体上对拟分析的数据进行宏观把握，为后续进行更深入的数据分析做好必要准备。

17.4.1 Stata 分析过程

描述性分析的步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入命令：

```
summarize  V2-V23,detail
```

03 设置完毕后，按键盘上的回车键，等待输出结果。

17.4.2 结果分析

在 Stata 14.0 主界面的结果窗口可以看到如图 17.2~图 17.12 所示的分析结果。

年底总人口

	Percentiles	Smallest		
1%	62	62		
5%	149	149		
10%	167	153	Obs	36
25%	337	167	Sum of Wgt.	36
50%	623.5		Mean	667.9444
		Largest	Std. Dev.	542.749
75%	768.5	1112		
90%	1112	1213	Variance	294576.5
95%	1379	1379	Skewness	2.99197
99%	3235	3235	Kurtosis	15.02719

地区生产总值

	Percentiles	Smallest		
1%	1219100	1219100		
5%	3424581	3424581		
10%	4086009	3936858	Obs	36
25%	1.18e+07	4086009	Sum of Wgt.	36
50%	2.28e+07		Mean	2.84e+07
		Largest	Std. Dev.	2.58e+07
75%	3.38e+07	6.80e+07		
90%	6.80e+07	7.11e+07	Variance	6.66e+14
95%	9.35e+07	9.35e+07	Skewness	1.895145
99%	1.22e+08	1.22e+08	Kurtosis	6.817154

图 17.2　V2 和 V3 描述性分析结果图

第一产业增加值

	Percentiles	Smallest		
1%	69412	69412		
5%	74200	74200		
10%	162659	149349	Obs	36
25%	361989.5	162659	Sum of Wgt.	36
50%	1015500		Mean	1248159
		Largest	Std. Dev.	1038455
75%	1648113	2492933		
90%	2492933	2757361	Variance	1.08e+12
95%	3476764	3476764	Skewness	1.36616
99%	4823900	4823900	Kurtosis	5.336955

第二产业增加值

	Percentiles	Smallest		
1%	312100	312100		
5%	1111209	1111209		
10%	2020998	1772896	Obs	36
25%	5291018	2020998	Sum of Wgt.	36
50%	1.02e+07		Mean	1.30e+07
		Largest	Std. Dev.	1.13e+07
75%	1.75e+07	2.81e+07		
90%	2.81e+07	2.89e+07	Variance	1.27e+14
95%	3.40e+07	3.40e+07	Skewness	1.853984
99%	5.68e+07	5.68e+07	Kurtosis	7.540296

图 17.3　V4 和 V5 描述性分析结果图

第三产业增加值

	Percentiles	Smallest		
1%	832800	832800		
5%	1502336	1502336		
10%	2557511	1820621	Obs	36
25%	5707481	2557511	Sum of Wgt.	36
50%	9936043		Mean	1.42e+07
		Largest	Std. Dev.	1.52e+07
75%	1.59e+07	3.39e+07		
90%	3.39e+07	4.15e+07	Variance	2.32e+14
95%	6.41e+07	6.41e+07	Skewness	2.362122
99%	6.74e+07	6.74e+07	Kurtosis	8.265398

客运量

	Percentiles	Smallest		
1%	2871	2871		
5%	2926	2926		
10%	3345	3303	Obs	36
25%	6426.5	3345	Sum of Wgt.	36
50%	12182.5		Mean	16043.19
		Largest	Std. Dev.	15179.27
75%	18740.5	30693		
90%	30693	43317	Variance	2.30e+08
95%	51180	51180	Skewness	2.363536
99%	77727	77727	Kurtosis	9.327829

图 17.4　V6 和 V7 描述性分析结果图

货运量

	Percentiles	Smallest		
1%	32	32		
5%	2441	2441		
10%	4582	4088	Obs	36
25%	7581	4582	Sum of Wgt.	36
50%	13304		Mean	18817.97
		Largest	Std. Dev.	16370.92
75%	22560.5	45142		
90%	45142	49970	Variance	2.68e+08
95%	50462	50462	Skewness	1.789018
99%	78108	78108	Kurtosis	6.337854

地方财政预算内收入

	Percentiles	Smallest		
1%	53800	53800		
5%	184123	184123		
10%	448495	276384	Obs	36
25%	815676.5	448495	Sum of Wgt.	36
50%	1631684		Mean	2860951
		Largest	Std. Dev.	4066302
75%	3109628	5404390		
90%	5404390	6580554	Variance	1.65e+13
95%	1.49e+07	1.49e+07	Skewness	3.182282
99%	2.07e+07	2.07e+07	Kurtosis	13.34594

图 17.5　V8 和 V9 描述性分析结果图

地方财政预算内支出

	Percentiles	Smallest		
1%	246227	246227		
5%	432058	432058		
10%	484708	450463	Obs	36
25%	1174302	484708	Sum of Wgt.	36
50%	1901096		Mean	3418107
		Largest	Std. Dev.	4358640
75%	3437557	7279563		
90%	7279563	7683886	Variance	1.90e+13
95%	1.65e+07	1.65e+07	Skewness	2.942701
99%	2.18e+07	2.18e+07	Kurtosis	11.82385

固定资产投资总额

	Percentiles	Smallest		
1%	876200	876200		
5%	1709130	1709130		
10%	2816095	1818297	Obs	36
25%	5932940	2816095	Sum of Wgt.	36
50%	1.35e+07		Mean	1.39e+07
		Largest	Std. Dev.	1.00e+07
75%	1.80e+07	2.39e+07		
90%	2.39e+07	3.16e+07	Variance	1.01e+14
95%	3.97e+07	3.97e+07	Skewness	1.215752
99%	4.46e+07	4.46e+07	Kurtosis	4.599039

图 17.6　V10 和 V11 描述性分析结果图

城乡居民储蓄年末余额

	Percentiles	Smallest		
1%	904700	904700		
5%	2970909	2970909		
10%	4309529	3264143	Obs	36
25%	6971610	4309529	Sum of Wgt.	36
50%	1.46e+07		Mean	1.99e+07
		Largest	Std. Dev.	2.11e+07
75%	1.98e+07	3.79e+07		
90%	3.79e+07	5.86e+07	Variance	4.45e+14
95%	9.11e+07	9.11e+07	Skewness	2.412383
99%	9.33e+07	9.33e+07	Kurtosis	8.556773

在岗职工平均工资

	Percentiles	Smallest		
1%	19992	19992		
5%	21019	21019		
10%	22156	22104	Obs	36
25%	23918.5	22156	Sum of Wgt.	36
50%	26630.5		Mean	28881.22
		Largest	Std. Dev.	7564.119
75%	31017	40561		
90%	40561	46508	Variance	5.72e+07
95%	46954	46954	Skewness	1.349985
99%	49311	49311	Kurtosis	3.911172

图 17.7　V12 和 V13 描述性分析结果图

年末邮政局数

	Percentiles	Smallest		
1%	36	36		
5%	76	76		
10%	110	95	Obs	36
25%	178	110	Sum of Wgt.	36
50%	248.5		Mean	447.5556
		Largest	Std. Dev.	628.198
75%	401.5	842		
90%	842	1027	Variance	394632.8
95%	1981	1981	Skewness	3.621824
99%	3468	3468	Kurtosis	16.80621

年末固定电话用户数

	Percentiles	Smallest		
1%	31	31		
5%	54	54		
10%	93	61	Obs	36
25%	172.5	93	Sum of Wgt.	36
50%	285		Mean	318.1111
		Largest	Std. Dev.	221.5589
75%	397.5	643		
90%	643	723	Variance	49088.33
95%	915	915	Skewness	1.486064
99%	1022	1022	Kurtosis	5.324049

图 17.8　V14 和 V15 描述性分析结果图

社会商品零售总额

	Percentiles	Smallest		
1%	558000	558000		
5%	1281040	1281040		
10%	1893831	1389869	Obs	36
25%	4287444	1893831	Sum of Wgt.	36
50%	9598571		Mean	1.05e+07
		Largest	Std. Dev.	8835481
75%	1.33e+07	1.92e+07		
90%	1.92e+07	2.60e+07	Variance	7.81e+13
95%	3.80e+07	3.80e+07	Skewness	1.757778
99%	3.85e+07	3.85e+07	Kurtosis	6.256883

货物进出口总额

	Percentiles	Smallest		
1%	21908	21908		
5%	56740	56740		
10%	93952	71500	Obs	36
25%	308872	93952	Sum of Wgt.	36
50%	646591.5		Mean	3585619
		Largest	Std. Dev.	7106723
75%	3926126	7349356		
90%	7349356	1.93e+07	Variance	5.05e+13
95%	2.83e+07	2.83e+07	Skewness	2.764245
99%	2.88e+07	2.88e+07	Kurtosis	9.653569

图 17.9　V16 和 V17 描述性分析结果图

年末实有公共汽车营运车辆数

	Percentiles	Smallest		
1%	702	702		
5%	865	865		
10%	1353	992	Obs	36
25%	2321	1353	Sum of Wgt.	36
50%	3895.5		Mean	4835
		Largest	Std. Dev.	4069.403
75%	5635.5	9314		
90%	9314	10734	Variance	1.66e+07
95%	16944	16944	Skewness	2.066515
99%	19395	19395	Kurtosis	7.469046

影剧院数

	Percentiles	Smallest		
1%	4	4		
5%	5	5		
10%	5	5	Obs	36
25%	8.5	5	Sum of Wgt.	36
50%	17		Mean	19121
		Largest	Std. Dev.	114513.3
75%	59	114		
90%	114	150	Variance	1.31e+10
95%	153	153	Skewness	5.747048
99%	687115	687115	Kurtosis	34.02856

图 17.10　V18 和 V19 描述性分析结果图

普通高等学校在校学生数

	Percentiles	Smallest		
1%	265	265		
5%	12163	12163		
10%	52657	37665	Obs	36
25%	145546	52657	Sum of Wgt.	36
50%	297003.5		Mean	307963.1
		Largest	Std. Dev.	201054.7
75%	467697.5	570794		
90%	570794	624403	Variance	4.04e+10
95%	679924	679924	Skewness	.3916176
99%	778368	778368	Kurtosis	2.372773

医院数

	Percentiles	Smallest		
1%	53	53		
5%	75	75		
10%	108	101	Obs	36
25%	211	108	Sum of Wgt.	36
50%	267		Mean	1150.061
		Largest	Std. Dev.	4791.931
75%	460.5	686		
90%	686	1162	Variance	2.30e+07
95%	1447	1447	Skewness	5.716361
99%	29056	29056	Kurtosis	33.79682

图 17.11　V20 和 V21 描述性分析结果图

执业医师

	Percentiles	Smallest		
1%	1050	1050		
5%	3410	3410		
10%	4541	4368	Obs	36
25%	9462.5	4541	Sum of Wgt.	36
50%	15218		Mean	57140.67
		Largest	Std. Dev.	244875.4
75%	19242.5	38739		
90%	38739	48825	Variance	6.00e+10
95%	54989	54989	Skewness	5.727096
99%	1484003	1484003	Kurtosis	33.87881

环境污染治理投资总额

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	35
25%	12326	0	Sum of Wgt.	35
50%	72382		Mean	298009.8
		Largest	Std. Dev.	652770
75%	272202	653008		
90%	653008	996994	Variance	4.26e+11
95%	1217394	1217394	Skewness	4.165502
99%	3661231	3661231	Kurtosis	21.51808

图 17.12　V22 和 V23 描述性分析结果图

在如图 17.2~图 17.12 所示的分析结果中，可以得到很多信息。此处限于篇幅不再针对各个变量一一展开说明，以变量环境污染治理投资总额为例进行解释。

- 百分位数（Percentiles）：可以看出变量 V23 的第 1 个四分位数（25%）是 12326，第 2 个四分位数（50%）是 72382。
- 4 个最小值（Smallest）：变量环境污染治理投资总额最小的 4 个数据值分别是 0、0、0、0。
- 4 个最大值（Largest）：变量环境污染治理投资总额最大的 4 个数据值分别是 653008、996994、1217394、3661231。
- 平均值（Mean）和标准差（Std. Dev）：变量环境污染治理投资总额的平均值为 298009.8，

标准差是 652770。

- 偏度（Skewness）和峰度（Kurtosis）：变量环境污染治理投资总额的偏度为 4.165502，为正偏度。变量环境污染治理投资总额的峰度为 21.51808，有一个比正态分布更长的尾巴。

从上面的描述性分析结果中，可以比较轻松地看出，所有数据中没有极端数据，数据间的量纲差距也在可接受范围之内，可以进入下一步的分析过程。

17.5 相关分析

对于相关分析，我们准备进行以下几个部分：

- 对“地区生产总值”的 3 个组成部分（“第一产业增加值”“第二产业增加值”“第三产业增加值”）进行简单相关分析。
- 对“客运量”和“货运量”进行简单相关分析。
- 对“地方财政预算内收入”和“地方财政预算内支出”进行简单相关分析。
- 对“年底总人口”“地区生产总值”“环境污染治理投资总额”这 3 个变量进行简单相关分析。

1. 对“地区生产总值”的 3 个组成部分进行简单相关分析

操作步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入如下命令。

- correlate V4 V5 V6：本命令旨在使用简单相关分析方法研究“第一产业增加值”“第二产业增加值”“第三产业增加值”3 个变量之间的相关关系。
- pwcorr V4 V5 V6,sidak sig star(0.01)：本命令旨在判断“第一产业增加值”“第二产业增加值”“第三产业增加值”3 个变量之间的相关性在置信水平为 99%时是否显著。

03 设置完毕后，按键盘上的回车键，等待输出结果。

结果分析如图 17.13 和图 17.14 所示。从图 17.13 可以看出，只有“第二产业增加值”与“第三产业增加值”之间具有比较大的相关系数。

```
. correlate  V4 V5 V6
(obs=36)

             |     V4       V5       V6
-------------+---------------------------
          V4 |  1.0000
          V5 |  0.2088   1.0000
          V6 |  0.1208   0.8673   1.0000
```

图 17.13 相关分析结果图 1

```
. pwcorr V4 V5 V6,sidak sig star(0.01)

             |     V4       V5       V6
-------------+---------------------------
          V4 |  1.0000
             |
             |
          V5 |  0.2088   1.0000
             |  0.5284
             |
          V6 |  0.1208   0.8673*  1.0000
             |  0.8615   0.0000
```

图 17.14 相关分析结果图 2

从图 17.14 中可以看出，只有“第二产业增加值”与“第三产业增加值”之间具有很强的相关性，并且在 0.01 的显著性水平上显著，其他的变量之间相关性很不显著。

2. 对“客运量”和“货运量”进行简单相关分析

操作步骤如下：

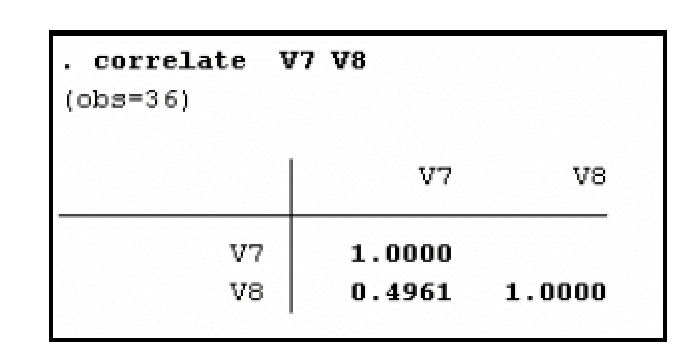
01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入如下命令。

- correlate V7 V8：本命令旨在使用简单相关分析方法研究“客运量”和“货运量”这 2 个变量之间的相关关系。
- pwcorr V7 V8,sidak sig star(0.01)：本命令旨在判断“客运量”和“货运量”这 2 个变量之间的相关性在置信水平为 99%时是否显著。

03 设置完毕后，按键盘上的回车键，等待输出结果。

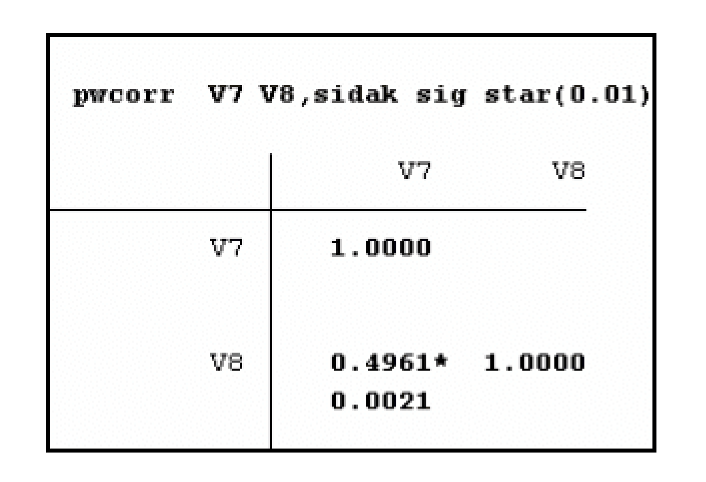
结果分析如图 17.15 和图 17.16 所示。从图 17.15 可以看出，“客运量”与“货运量”之间的相关系数不是很大。



```
. correlate V7 V8
(obs=36)

             |     V7      V8
-------------+-----------------
          V7 | 1.0000
          V8 | 0.4961  1.0000
```

图 17.15　相关分析结果图 3



```
pwcorr V7 V8,sidak sig star(0.01)

             |     V7      V8
-------------+-----------------
          V7 | 1.0000
             |
          V8 | 0.4961*  1.0000
             | 0.0021
```

图 17.16　相关分析结果图 4

从图 17.16 中可以看出，“客运量”与“货运量”之间虽然相关系数不是很大，但是这种相关性却很强，在 0.01 的显著性水平上显著。

3. 对“地方财政预算内收入”和“地方财政预算内支出”进行简单相关分析

操作步骤如下：

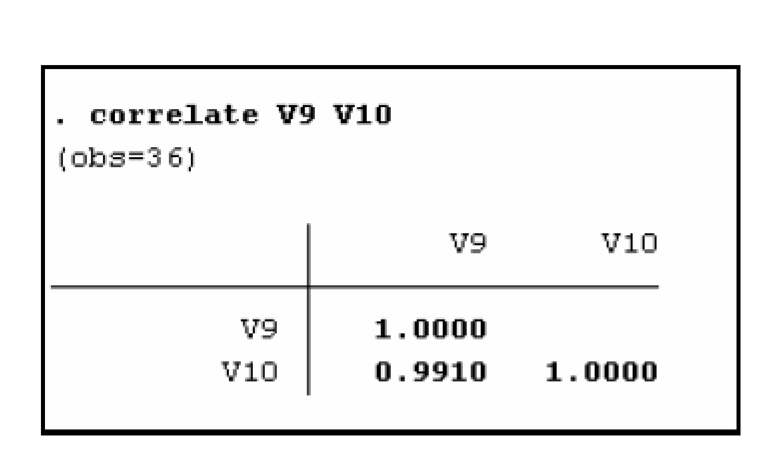
01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入如下命令。

- correlate V9 V10：本命令旨在使用简单相关分析方法研究“客运量”和“货运量”这 2 个变量之间的相关关系。
- pwcorr V9 V10,sidak sig star(0.01)：本命令旨在判断“客运量”和“货运量”这 2 个变量之间的相关性在置信水平为 99%时是否显著。

03 设置完毕后，按键盘上的回车键，等待输出结果。

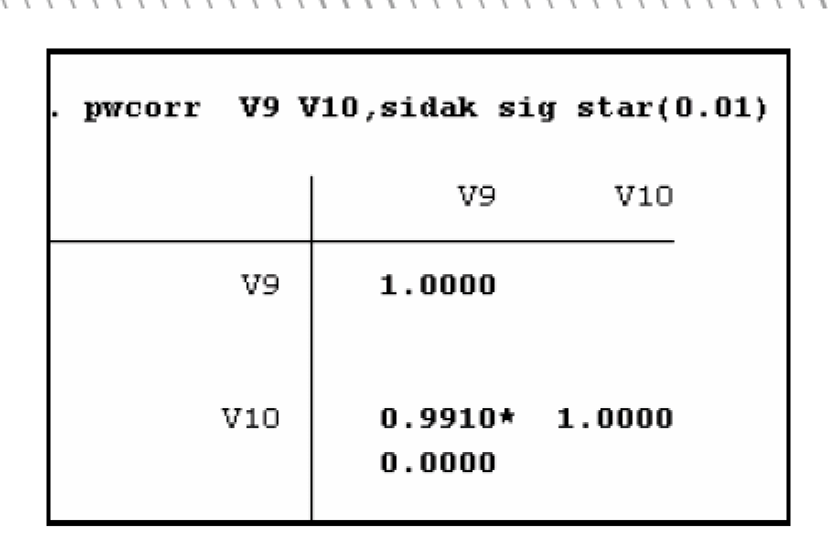
结果分析如图 17.17 和图 17.18 所示。从图 17.17 可以看出，“地方财政预算内收入”和“地方财政预算内支出”之间的相关系数很大。



```
. correlate V9 V10
(obs=36)

             |      V9      V10
-------------+------------------
          V9 |  1.0000
         V10 |  0.9910   1.0000
```

图 17.17　相关分析结果图 5



```
. pwcorr  V9 V10,sidak sig star(0.01)

             |      V9      V10
-------------+------------------
          V9 |  1.0000
             |
             |
         V10 |  0.9910*  1.0000
             |  0.0000
```

图 17.18　相关分析结果图 6

从图 17.18 中可以看出，“地方财政预算内收入”和“地方财政预算内支出”相关系数不是很大，而且这种相关性很强，在 0.01 的显著性水平上显著。

4. 对“年底总人口”“地区生产总值”“环境污染治理投资总额”进行简单相关分析

操作步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

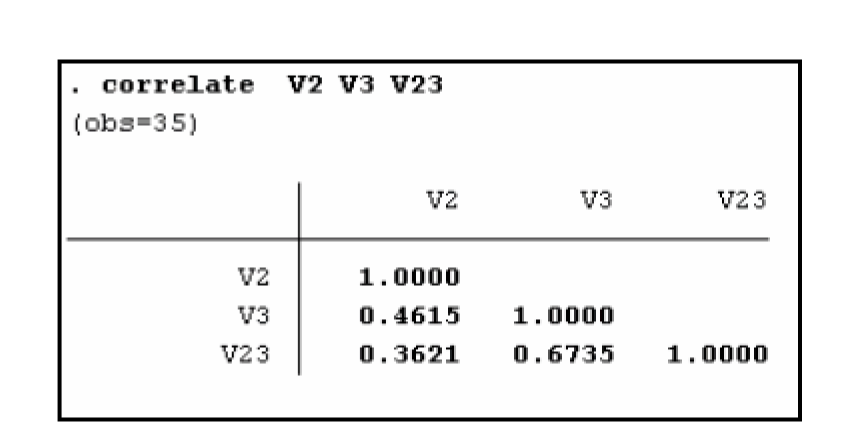
02 在主界面的“Command”文本框中输入如下命令。

- correlate V2 V3 V23：本命令旨在使用简单相关分析方法研究“年底总人口”“地区生产总值”“环境污染治理投资总额” 3 个变量之间的相关关系。
- pwcorr V2 V3 V23,sidak sig star(0.01)：本命令旨在判断“年底总人口”“地区生产总值”“环境污染治理投资总额”3 个变量之间的相关性在置信水平为 99%时是否显著。

03 设置完毕后，按键盘上的回车键，等待输出结果。

结果分析如图 17.19 和图 17.20 所示。从图 17.19 可以看出，年底总人口与地区生产总值为正相关但相关系数不大；年底总人口与环境污染治理投资总额之间也为正相关，而且相关系数也不大；地区生产总值与环境污染治理投资总额之间也为正相关，相关系数较大。

从图 17.20 中可以看出，只有地区生产总值与环境污染治理投资总额之间的相关关系非常显著（在 0.01 的水平上显著）。



```
. correlate  V2 V3 V23
(obs=35)

             |      V2       V3      V23
-------------+---------------------------
          V2 |  1.0000
          V3 |  0.4615   1.0000
         V23 |  0.3621   0.6735   1.0000
```

图 17.19　相关分析结果图 7

```
. pwcorr  V2 V3 V23,sidak sig star(0.01)

             |      V2       V3      V23
-------------+---------------------------
          V2 |  1.0000
             |
             |
          V3 |  0.4685   1.0000
             |  0.0118
             |
         V23 |  0.3621   0.6735*  1.0000
             |  0.0945   0.0000
```

图 17.20　相关分析结果图 8

17.6　回归分析

对于回归分析，我们准备以“地区生产总值”为因变量，以“年底总人口”“客运量”“货运量”“地方财政预算内收入”“地方财政预算内支出”“固定资产投资总额”“城乡居

民储蓄年末余额”“在岗职工平均工资”“年末邮政局数”“年末固定电话用户数”“社会商品零售总额”“货物进出口总额”“年末实有公共汽车营运车辆数”“影剧院数”“普通高等学校在校学生数”“医院数”“执业医师”“环境污染治理投资总额”等为自变量，进行多重线性回归。

建立线性模型：

地区生产总值= a*年底总人口+b*客运量+c*货运量+d*地方财政预算内收入+e*地方财政预算内支出+f*固定资产投资总额+g*城乡居民储蓄年末余额+h*在岗职工平均工资+i*年末邮政局数+j*年末固定电话用户数+k*社会商品零售总额+l*货物进出口总额+m*年末实有公共汽车营运车辆数+n*影剧院数+o*普通高等学校在校学生数+p*医院数+q*执业医师+r*环境污染治理投资总额+ u

普通最小二乘回归分析的步骤及结果如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入如下命令。

- sw regress V3 V2 V7-V23,pr(0.10)：本命令的含义是使用逐步回归分析方法，以“地区生产总值”为因变量，以“年底总人口”“客运量”“货运量”“地方财政预算内收入”“地方财政预算内支出”“固定资产投资总额”“城乡居民储蓄年末余额”“在岗职工平均工资”“年末邮政局数”“年末固定电话用户数”“社会商品零售总额”“货物进出口总额”“年末实有公共汽车营运车辆数”“影剧院数”“普通高等学校在校学生数”“医院数”“执业医师”“环境污染治理投资总额”等为自变量，进行多重线性回归。
- predict yhat：本命令旨在获得因变量的拟合值。
- predict e,resid：本命令旨在获得回归模型的估计残差。
- rvfplot：本命令旨在绘制残差与回归得到的拟合值的散点图，探索数据是否存在异方差。
- estat imtest,white：本命令为怀特检验，旨在检验数据是否存在异方差。
- estat hettest,iid：本命令为 BP 检验，旨在使用得到的拟合值来检验数据是否存在异方差。
- estat hettest,rhs iid：本命令为 BP 检验，旨在使用方程右边的解释数据来检验变量是否存在异方差。

03 设置完毕后，按键盘上的回车键进行确认。

在 Stata 14.0 主界面的结果窗口可以看到如图 17.21~图 17.27 所示的分析结果。

图 17.21 是使用逐步回归分析方法，以“地区生产总值”为因变量，以“年底总人口”“客运量”“货运量”“地方财政预算内收入”“地方财政预算内支出”“固定资产投资总额”“城乡居民储蓄年末余额”“在岗职工平均工资”“年末邮政局数”“年末固定电话用户数”“社会商品零售总额”“货物进出口总额”“年末实有公共汽车营运车辆数”“影剧院数”“普通高等学校在校学生数”“医院数”“执业医师”“环境污染治理投资总额”等为自变量，进行多重线性回归的结果。

```
. sw regress  V3 V2 V7-V23,pr(0.10)
                        begin with full model
p = 0.8172 >= 0.1000  removing V13
p = 0.6659 >= 0.1000  removing V23
p = 0.6435 >= 0.1000  removing V15
p = 0.5136 >= 0.1000  removing V20
p = 0.4071 >= 0.1000  removing V2
p = 0.5773 >= 0.1000  removing V10
p = 0.2603 >= 0.1000  removing V9
p = 0.2561 >= 0.1000  removing V22
p = 0.3142 >= 0.1000  removing V12
p = 0.2027 >= 0.1000  removing V7

      Source |       SS       df       MS              Number of obs =      35
-------------+------------------------------           F(  8,    26) =  651.47
       Model |  2.2732e+16     8  2.8415e+15           Prob > F      =  0.0000
    Residual |  1.1340e+14    26  4.3617e+12           R-squared     =  0.9950
-------------+------------------------------           Adj R-squared =  0.9935
       Total |  2.2846e+16    34  6.7193e+14           Root MSE      =  2.1e+06

------------------------------------------------------------------------------
          V3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         V19 |  -141.4416   72.19314    -1.96   0.061    -289.8367    6.953492
         V14 |  -4391.273   1535.017    -2.86   0.008    -7546.545       -1236
          V8 |   196.4853     41.527     4.73   0.000     111.1254    281.8453
         V17 |   1.320081   .1056014    12.50   0.000     1.103014    1.537147
         V18 |  -749.7279   342.2635    -2.19   0.038     -1453.26   -46.19524
         V11 |   .3700798   .1297517     2.85   0.008     .1033713    .6367882
         V21 |   4141.523   1796.039     2.31   0.029     449.7115    7833.335
         V16 |   1.701424   .2203774     7.72   0.000     1.248432    2.154417
       _cons |   534506.5   679459.6     0.79   0.439    -862142.7     1931156
------------------------------------------------------------------------------
```

图 17.21　回归分析结果图 1

从上述分析结果中可以看出共有 35 个样本参与了分析，模型的 F 值(8, 26) =651.47，P 值（Prob > F） = 0.0000，说明模型整体上是非常显著的。模型的可决系数（R-squared）为 0.9950，模型修正的可决系数（Adj R-squared）为 0.9935，说明模型的解释能力是非常优秀且接近完美的。

模型经过 10 次剔除变量后得到最终结果。第 1 个模型是包含全部自变量的全模型，该模型中 V13 变量的系数显著性 P 值高达 0.8172，被剔除掉；第 2 个模型是包含全部自变量的全模型，该模型中 V23 变量的系数显著性 P 值高达 0.6659，被剔除掉；第 3 个模型是包含全部自变量的全模型，该模型中 V15 变量的系数显著性 P 值高达 0.6435，被剔除掉；第 4 个模型是包含全部自变量的全模型，该模型中 V20 变量的系数显著性 P 值高达 0.5136，被剔除掉；第 5 个模型是包含全部自变量的全模型，该模型中 V2 变量的系数显著性 P 值高达 0.4071，被剔除掉；第 6 个模型是包含全部自变量的全模型，该模型中 V10 变量的系数显著性 P 值高达 0.5773，被剔除掉；第 7 个模型是包含全部自变量的全模型，该模型中 V9 变量的系数显著性 P 值高达 0.2603，被剔除掉；第 8 个模型是包含全部自变量的全模型，该模型中 V22 变量的系数显著性 P 值高达 0.2561，被剔除掉；第 9 个模型是包含全部自变量的全模型，该模型中 V12 变量的系数显著性 P 值高达 0.3142，被剔除掉；第 10 个模型是包含全部自变量的全模型，该模型中 V7 变量的系数显著性 P 值高达 0.2027，被剔除掉；剔除掉上述自变量以后得到最终回归模型。

在最终回归模型中，变量 V19 的系数标准误是 72.19314，t 值为-1.96，P 值为 0.061，系数是比较显著的，95%的置信区间为[-289.8367, 6.953492]。变量 V14 的系数标准误是 1535.017，t 值为-2.86，P 值为 0.008，系数是非常显著的，95%的置信区间为[-7546.545, -1236]。变量 V8 的系数标准误是 41.527，t 值为 4.73，P 值为 0.000，系数是非常显著的，95%的置信区间为

[111.1254, 281.8453]。变量 V17 的系数标准误是 0.1056014，t 值为 12.50，P 值为 0.000，系数是非常显著的，95%的置信区间为[1.103014, 1.537147]。变量 V18 的系数标准误是 342.2635，t 值为-2.19，P 值为 0.038，系数是非常显著的，95%的置信区间为[-1453.26, -46.19524]。变量 V11 的系数标准误是 0.1297517，t 值为 2.85，P 值为 0.008，系数是非常显著的，95%的置信区间为[0.1033713, 0.6367882]。变量 V21 的系数标准误是 1796.039，t 值为 2.31，P 值为 0.029，系数是非常显著的，95%的置信区间为[449.7115, 7833.335]。变量 V16 的系数标准误是 0.2203774，t 值为 7.72，P 值为 0.000，系数是非常显著的，95%的置信区间为[1.248432, 2.154417]。常数项的系数标准误是 679459.6，t 值为 0.79，P 值为 0.439，系数是非常不显著的，95%的置信区间为[-862142.7, 1931156]。

最终最小二乘回归模型的方程是：

地区生产总值= 196.4853*货运量+0.3700798*固定资产投资总额-4391.273*年末邮政局数+1.701424*社会商品零售总额+ 1.320081*货物进出口总额-749.7279 *年末实有公共汽车营运车辆数-141.4416*影剧院数+ 4141.523*医院数+534506.5

图 17.22 是对因变量的拟合值的预测。

	V16	V17	V18	V19	V20	V21	V22	V23	yhat
1	38002053	19294630	19395	153	567875	686	54989	493300	9.41e+07
2	16037390	7154981	7489	28	371136	411	26228	1217394	4.84e+07
3	8210983	512981	2024	18	316796	394	18773	0	2.19e+07
4	5159090	810678	1769	16	298188	273	13646	234307	1.55e+07
5	4307960	93952	1353	7	164998	146	6160	29348	1.08e+07
6	12318501	606943	5096	58	317450	314	19700	141618	3.12e+07
7	9832589	3874480	4784	5	219982	236	14119	191069	3.18e+07
8	7783494	694434	3788	17	331019	320	15670	0	1.93e+07
9	10359652	299805	4705	69	384888	472	18000	341754	2.10e+07
10	38477918	28297305	16944	150	484873	534	48825	3661231	1.22e+08
11	13804635	3619974	6709	12	679924	167	15705	996994	3.42e+07
12	12963123	4342666	5435	71	366160	1162	20701	0	3.83e+07
13	10354628	5649909	2777	21	126094	266	15418	224191	3.34e+07
14	4690023	624839	2409	17	295819	225	8539	43819	1.40e+07
15	9409928	1864051	2270	114	233133	221	11601	74600	2.33e+07
16	3620462	3977772	2543	67	105545	53	5950	272202	1.42e+07
17	4266928	319472	2448	9	481107	166	6363	61164	1.04e+07
18	11031402	621804	4003	11	570794	243	15800	0	2.48e+07
19	11991772	4572534	4524	40	264917	251	15018	653008	3.70e+07
20	9787214	317939	3518	14	495719	307	16002	0	2.24e+07
21	15183023	996179	6600	60	778368	227	21541	72382	3.32e+07
22	10370277	407151	3252	8	454288	265	14683	158808	2.14e+07
23	25950035	7349356	9314	687115	265	29056	1484003	0	7.11e+07
24	19150276	28753345	10734	56	58910	101	18785	12534	6.83e+07
25	5156225	128596	2817	10	238375	201	11879	261235	1.12e+07
26	1893831	193390	865	6	91152	75	4368	16831	5161705
27	16612275	744546	8411	69	413655	1447	38739	619723	4.23e+07
28	13572003	951552	5158	17	540626	595	28673	0	3.62e+07
29	2793885	169227	2372	5	209499	268	10035	.	7264870

图 17.22 回归分析结果图 2

因变量预测拟合值是根据自变量的值和得到的回归方程计算出来的，主要用于预测未来。在图 17.22 中，可以看到 yhat 的值与 var3 的值是比较相近的，所以拟合的回归模型还是不错的。

图 17.23 是回归分析得到的残差序列。

	V17	V18	V19	V20	V21	V22	V23	yhat	e
1	19294630	19395	153	567875	686	54989	493300	9.41e+07	-575593.6
2	7154981	7489	28	371136	411	26228	1217394	4.84e+07	2096997
3	512981	2024	18	316796	394	18773	0	2.19e+07	1726529
4	810678	1769	16	298188	273	13646	234307	1.55e+07	-2902802
5	93952	1353	7	164998	146	6160	29348	1.08e+07	163034.9
6	606943	5096	58	317450	314	19700	141618	3.12e+07	1055269
7	3874480	4784	5	219982	236	14119	191069	3.18e+07	-507858.6
8	694434	3788	17	331019	320	15670	0	1.93e+07	1607870
9	299805	4705	69	384888	472	18000	341754	2.10e+07	3325525
10	28297305	16944	150	484873	534	48825	3661231	1.22e+08	-820.311
11	3619974	6709	12	679924	167	15705	996994	3.42e+07	-1376059
12	4342666	5435	71	366160	1162	20701	0	3.83e+07	2750722
13	5649909	2777	21	126094	266	15418	224191	3.34e+07	969433.4
14	624839	2409	17	295819	225	8539	43819	1.40e+07	-657705.8
15	1864051	2270	114	233133	221	11601	74600	2.33e+07	-3513743
16	3977772	2543	67	105545	53	5950	272202	1.42e+07	-286734.7
17	319472	2448	9	481107	166	6363	61164	1.04e+07	3450905
18	621804	4003	11	570794	243	15800	0	2.48e+07	829734.2
19	4572534	4524	40	264917	251	15018	653008	3.70e+07	848750.6
20	317939	3518	14	495719	307	16002	0	2.24e+07	2472236
21	996179	6600	60	778368	227	21541	72382	3.32e+07	-1755588
22	407151	3252	8	454288	265	14683	158808	2.14e+07	509667.4
23	7349356	9314	687115	265	29056	1484003	0	7.11e+07	414.1629
24	28753345	10734	56	58910	101	18785	12534	6.83e+07	-264358.2
25	128596	2817	10	238375	201	11879	261235	1.12e+07	-494124.1
26	193390	865	6	91152	75	4368	16831	5161705	-1224846
27	744546	8411	69	413655	1447	38739	619723	4.23e+07	-1053245
28	951552	5158	17	540626	595	28673	0	3.62e+07	-2929076
29	169227	2372	5	209499	268	10035	.	7264870	-327699.7

图 17.23　回归分析结果图 3

图 17.24 是上面两步得到的残差与得到的拟合值的散点图。

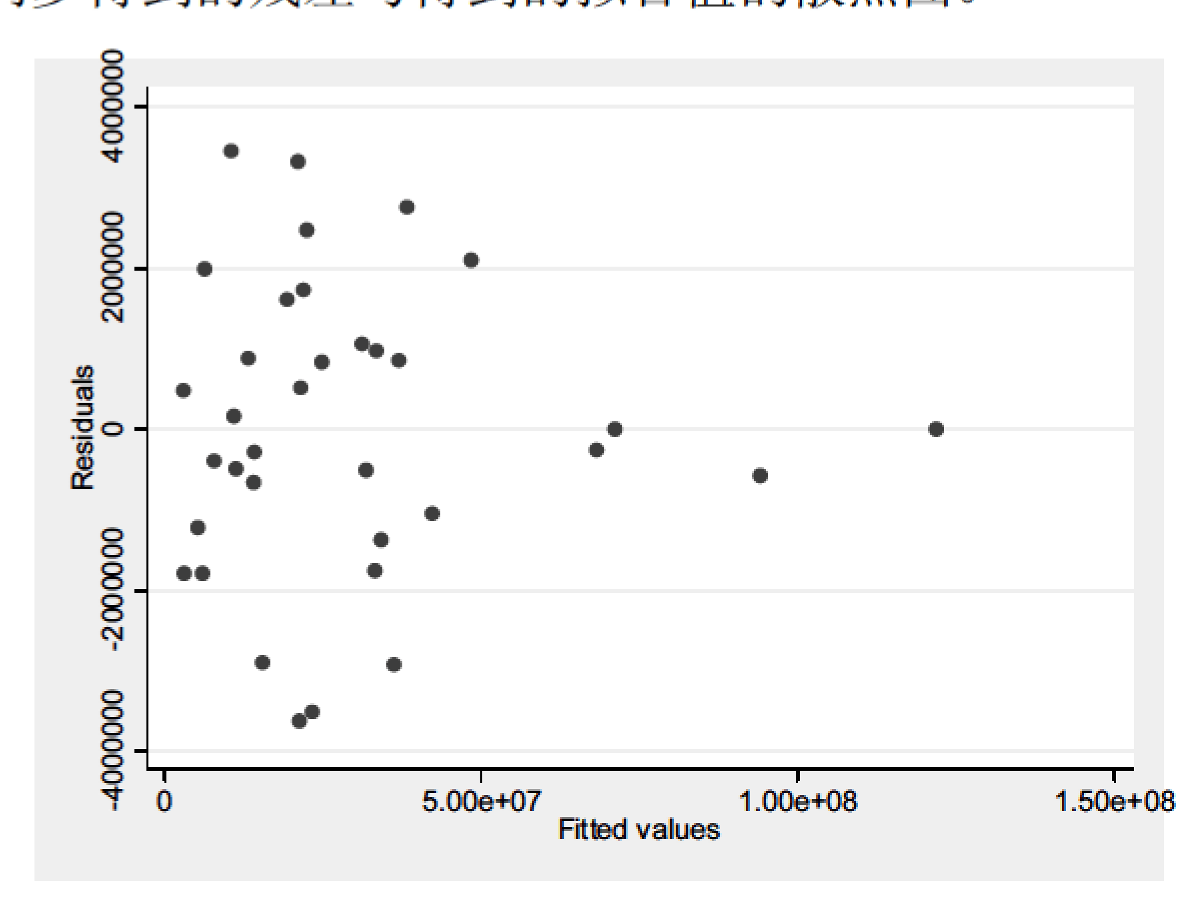


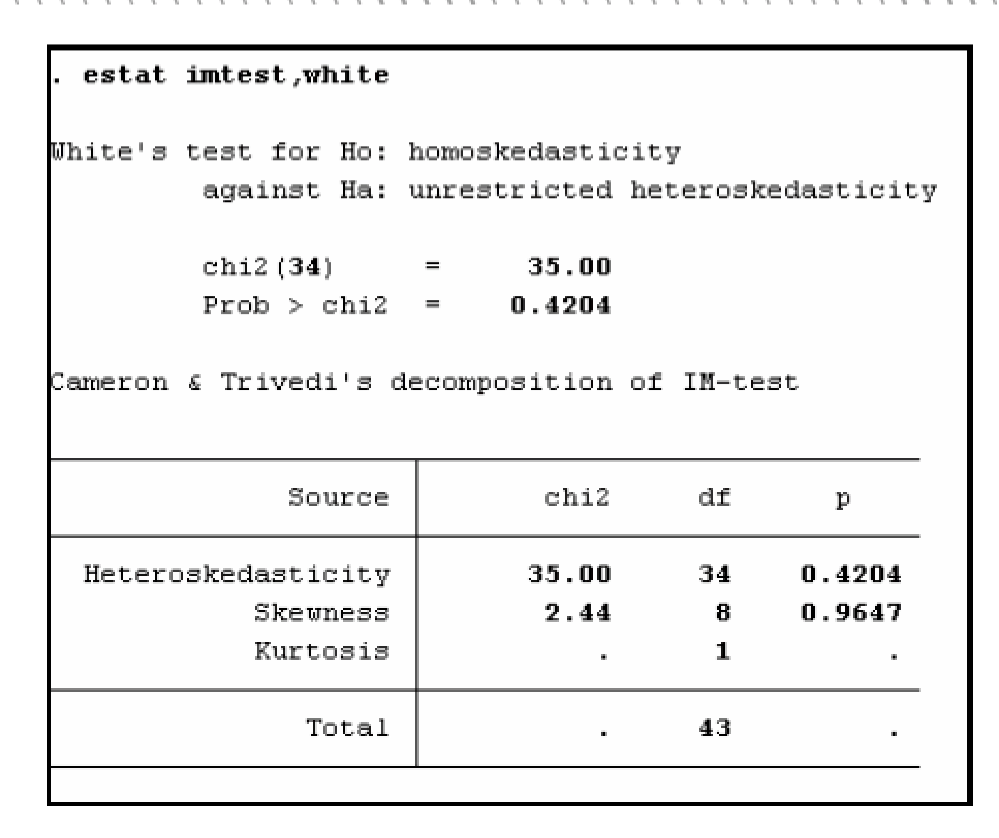
图 17.24　回归分析结果图 4

从图 17.24 中可以看出，残差并没有随着拟合值的大小的不同而不同，而是围绕 0 值上下随机波动，所以，数据很可能是不存在异方差的。

图 17.25 是怀特检验的检验结果。

怀特检验的原假设是数据为同方差。从图 17.25 中可以看出，P 值为 0.4204，非常显著地接受了同方差的原假设，认为不存在异方差。

图 17.26~图 17.27 是 BP 检验的检验结果。其中，图 17.26 是使用得到的拟合值对数据进行异方差检验的结果，图 17.27 是使用方程右边的解释变量对数据进行异方差检验的结果。



```
. estat imtest,white

White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(34)      =    35.00
         Prob > chi2   =   0.4204

Cameron & Trivedi's decomposition of IM-test

              Source |       chi2     df      p
---------------------+-----------------------------
  Heteroskedasticity |      35.00     34    0.4204
            Skewness |       2.44      8    0.9647
            Kurtosis |          .      1         .
---------------------+-----------------------------
               Total |          .     43         .
---------------------------------------------------
```

图 17.25 回归分析结果图 5

```
. estat hettest,iid

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of V3

         chi2(1)      =     1.44
         Prob > chi2  =   0.2309
```

图 17.26 回归分析结果图 6

```
. estat hettest,rhs iid

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: V19 V14 V8 V17 V18 V11 V21 V16

         chi2(8)      =     7.13
         Prob > chi2  =   0.5222
```

图 17.27 回归分析结果图 7

BP 检验的原假设是数据为同方差。从图 17.26 和图 17.27 中可以看出，P 值均大于 0.05，非常显著地接受了同方差的原假设，认为不存在异方差，所以我们没有必要使用稳健的标准差进行回归。

经过以上最小二乘回归分析，可以发现我国城市的地区生产总值与社会商品零售总额、货物进出口总额、货运量、固定资产投资总额、年末邮政局数、影剧院数、医院数、年末实有公共汽车营运车辆数有显著关系，与其他变量之间的关系并不显著。其中，固定资产投资总额、社会商品零售总额、货物进出口总额、医院数、货运量对地区生产总值起正向作用，尤其是医院数和货运量，每增加一个单位，地区生产总值就分别增加 4141.523 个单位和 196.4853 个单位，而年末邮政局数、影剧院数、年末实有公共汽车营运车辆数对地区生产总值起反向作用。

17.7 因子分析

对于因子分析，我们将对构成城市综合经济实力的各个变量提取公因子。

操作步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中分别输入如下命令并按键盘上的回车键进行确认。

- factor V2 V3 V7-V23,pcf: 本命令的含义是采用主成分因子法对构成城市综合经济实力的各个变量进行因子分析。
- rotate: 本命令的含义是采用最大方差正交旋转法对因子结构进行旋转。
- loadingplot,factors(2) yline(0) xline(0): 本命令的含义是绘制因子旋转后的因子载荷图。

- predict f1 f2 f3：本命令的含义是展示因子分析后各个样本的因子得分情况。
- correlate f1 f2 f3：本命令的含义是展示系统提取的 3 个主因子的相关系数矩阵。
- scoreplot,mlabel(V1) yline(0) xline(0)：本命令的含义是展示每个样本的因子得分示意图。
- estat kmo：本命令的含义是展示本例因子分析的 KMO 检验结果。
- screeplot：本命令的含义是展示本例因子分析所提取的各个因子的特征值碎石图。

03 设置完毕后，等待输出结果。

在 Stata 14.0 主界面的结果窗口可以看到如图 17.28~图 17.36 所示的分析结果。

图 17.28 展示的是因子分析的基本情况。

```
. factor  V2 V3 V7-V23,pcf
(obs=35)

Factor analysis/correlation                    Number of obs    =         35
    Method: principal-component factors        Retained factors =          3
    Rotation: (unrotated)                      Number of params =         54
```

Factor	Eigenvalue	Difference	Proportion	Cumulative
Factor1	10.84298	7.05651	0.5707	0.5707
Factor2	3.78647	1.71390	0.1993	0.7700
Factor3	2.07257	1.32814	0.1091	0.8791
Factor4	0.74443	0.13619	0.0392	0.9182
Factor5	0.60824	0.26697	0.0320	0.9502
Factor6	0.34128	0.14581	0.0180	0.9682
Factor7	0.19547	0.06271	0.0103	0.9785
Factor8	0.13276	0.03027	0.0070	0.9855
Factor9	0.10249	0.05017	0.0054	0.9909
Factor10	0.05231	0.01582	0.0028	0.9936
Factor11	0.03649	0.00277	0.0019	0.9956
Factor12	0.03373	0.00901	0.0018	0.9973
Factor13	0.02471	0.01054	0.0013	0.9986
Factor14	0.01418	0.00775	0.0007	0.9994
Factor15	0.00643	0.00270	0.0003	0.9997
Factor16	0.00373	0.00231	0.0002	0.9999
Factor17	0.00142	0.00112	0.0001	1.0000
Factor18	0.00030	0.00028	0.0000	1.0000
Factor19	0.00002	.	0.0000	1.0000

```
LR test: independent vs. saturated: chi2(171) = 1584.71 Prob>chi2 = 0.0000
```

Factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Factor3	Uniqueness
V2	0.5892	-0.0322	0.7383	0.1066
V3	0.9744	-0.0732	-0.1292	0.0284
V7	0.4814	0.4219	0.6115	0.2164
V8	0.8267	0.0213	0.2043	0.2744
V9	0.9285	-0.2710	-0.2089	0.0207
V10	0.9490	-0.2462	-0.1236	0.0235
V11	0.9055	-0.2150	0.2575	0.0675
V12	0.9705	-0.0423	-0.1218	0.0415
V13	0.6664	-0.0102	-0.5435	0.2604
V14	0.6002	0.7339	0.1857	0.0668
V15	0.9612	-0.0582	0.1802	0.0402
V16	0.9756	-0.0493	-0.0201	0.0454
V17	0.7710	-0.2275	-0.4619	0.1405
V18	0.9337	-0.1517	-0.0766	0.0994
V19	0.3384	0.9129	-0.1438	0.0315
V20	0.3135	-0.4137	0.5538	0.4239
V21	0.3625	0.9114	-0.1140	0.0250
V22	0.3790	0.8992	-0.1318	0.0305
V23	0.6874	-0.3959	-0.1236	0.3555

图 17.28　因子分析结果图 1

图 17.28 的上半部分说明的是因子分析模型的一般情况，从图中可以看出共有 35 个样本（Number of obs = 35）参与了分析，提取保留的因子共有 3 个（Retained factors = 3），模型

LR 检验的卡方值（LR test: independent vs. saturated: chi2(171)）为 1584.71，P 值（Prob>chi2）为 0.0000，模型非常显著。图 17.28 的上半部分最左列（Factor）说明的是因子名称，可以看出模型共提取了 19 个因子。Eigenvalue 列表示的是提取因子的特征值情况，只有前 3 个因子的特征值是大于 1 的，其中第 1 个因子的特征值是 10.84298，第 2 个因子的特征值是 3.78647。Proportion 列表示的是提取因子的方差贡献率，其中第 1 个因子的方差贡献率为 57.07%，第 2 个因子的方差贡献率为 19.93%。Cumulative 列表示的是提取因子的累计方差贡献率，其中前两个因子的累计方差贡献率为 77%。

图 17.28 的下半部分说明的是模型的因子载荷矩阵以及变量的未被解释部分。其中，Variable 列表示的是变量名称，Factor1、Factor2、Factor3 这 3 列分别说明的是提取的前 3 个主因子（特征值大于 1）对各个变量的解释程度，本例中，Factor1 主要解释的是 V2、V3、V7、V8、V9、V10~V18、V23 变量的信息，Factor2 主要解释的是 V7、V14、V19、V21、V22 变量的信息，Factor3 主要解释的是 V2、V7、V20 这 3 个变量的信息。Uniqueness 列表示变量未被提取的前 4 个主因子解释的部分，可以发现在舍弃其他主因子的情况下，信息的损失量是比较小的。

图 17.29 展示的是对因子结构进行旋转的结果。经过学者们的研究表明，旋转操作有助于进一步简化因子结构。Stata 14.0 支持的旋转方式有两种：一种是最大方差正交旋转，一般适用于相互独立的因子或者成分，也是系统默认的情况；另外一种是 Promax 斜交旋转，它允许因子或者成分之间存在相关关系。此处我们选择系统默认方式，当然我们后面的操作也证明了这种方式的恰当性。

```
. rotate

Factor analysis/correlation                    Number of obs    =        35
    Method: principal-component factors        Retained factors =         3
    Rotation: orthogonal varimax (Kaiser off)  Number of params =        54
```

Factor	Variance	Difference	Proportion	Cumulative
Factor1	9.34431	5.03419	0.4918	0.4918
Factor2	4.31012	1.26255	0.2268	0.7187
Factor3	3.04758	.	0.1604	0.8791

LR test: independent vs. saturated: chi2(171) = 1584.71 Prob>chi2 = 0.0000

Rotated factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Factor3	Uniqueness
V2	0.3063	0.0567	0.8924	0.1066
V3	0.9407	0.2133	0.2027	0.0284
V7	0.1151	0.4736	0.7390	0.2164
V8	0.6739	0.2284	0.4683	0.2744
V9	0.9831	0.0198	0.1114	0.0207
V10	0.9671	0.0404	0.1988	0.0235
V11	0.7979	0.0191	0.5437	0.0675
V12	0.9257	0.2409	0.2085	0.0415
V13	0.7760	0.2307	-0.2902	0.2604
V14	0.2651	0.8483	0.3786	0.0668
V15	0.8262	0.1922	0.4901	0.0402
V16	0.9001	0.2253	0.3061	0.0454
V17	0.9085	0.0437	-0.1795	0.1405
V18	0.9105	0.1216	0.2384	0.0994
V19	0.0809	0.9806	-0.0186	0.0315
V20	0.2290	-0.3652	0.6247	0.4239
V21	0.0936	0.9828	0.0176	0.0250
V22	0.1177	0.9775	0.0063	0.0305
V23	0.7756	-0.1750	0.1109	0.3555

Factor rotation matrix

	Factor1	Factor2	Factor3
Factor1	0.9011	0.2771	0.3334
Factor2	-0.2954	0.9553	0.0046
Factor3	-0.3173	-0.1027	0.9428

图 17.29　因子分析结果图 2

图 17.29 包括 3 部分内容，第 1 部分说明的是因子旋转模型的一般情况，从图中我们可以看出共有 35 个样本（Number of obs = 35）参与了分析，提取保留的因子共有 3 个（Retained factors = 3），模型 LR 检验的卡方值（LR test: independent vs. saturated: chi2(171)）为 1584.71，P 值（Prob>chi2）为 0.0000，模型非常显著。图 17.29 的上半部分最左列（Factor）说明的是因子名称，可以看出模型共保留了 19 个因子。Variance 列表示的是提取因子的特征值情况，只有前 3 个因子的特征值是大于 1 的，其中第 1 个因子的特征值是 9.34431，第 2 个因子的特征值是 4.31012。Proportion 列表示的是提取因子的方差贡献率，其中第 1 个因子的方差贡献率为 49.18%，第 2 个因子的方差贡献率为 22.68%。Cumulative 列表示的是提取因子的累计方差贡献率，其中前两个因子的累计方差贡献率为 71.87%。

图 17.29 的第 2 部分说明的是模型的因子载荷矩阵以及变量的未被解释部分。其中，Variable 列表示的是变量名称，Factor1、Factor2、Factor3 这 3 列分别说明的是旋转提取的 3 个主因子对各个变量的解释程度，本例中，Factor1 主要解释的是 V3、V8、V9、V10~V13、V15~V18、V23 变量的信息，Factor2 主要解释的是 V14、V19、V21、V22 变量的信息，Factor3 主要解释的是 V2、V7、V20 这 3 个变量的信息。Uniqueness 列表示变量未被提取的前 3 个主因子解释的部分，可以发现在舍弃其他主因子的情况下，信息的损失量是很小的。

图 17.29 的第 3 部分展示的是因子旋转矩阵的一般情况，提取的 4 个因子相关关系不明显。

图 17.30 展示的是因子旋转后的因子载荷图。因子载荷图可以使用户更加直观地看出各个变量被前两个因子的解释情况。

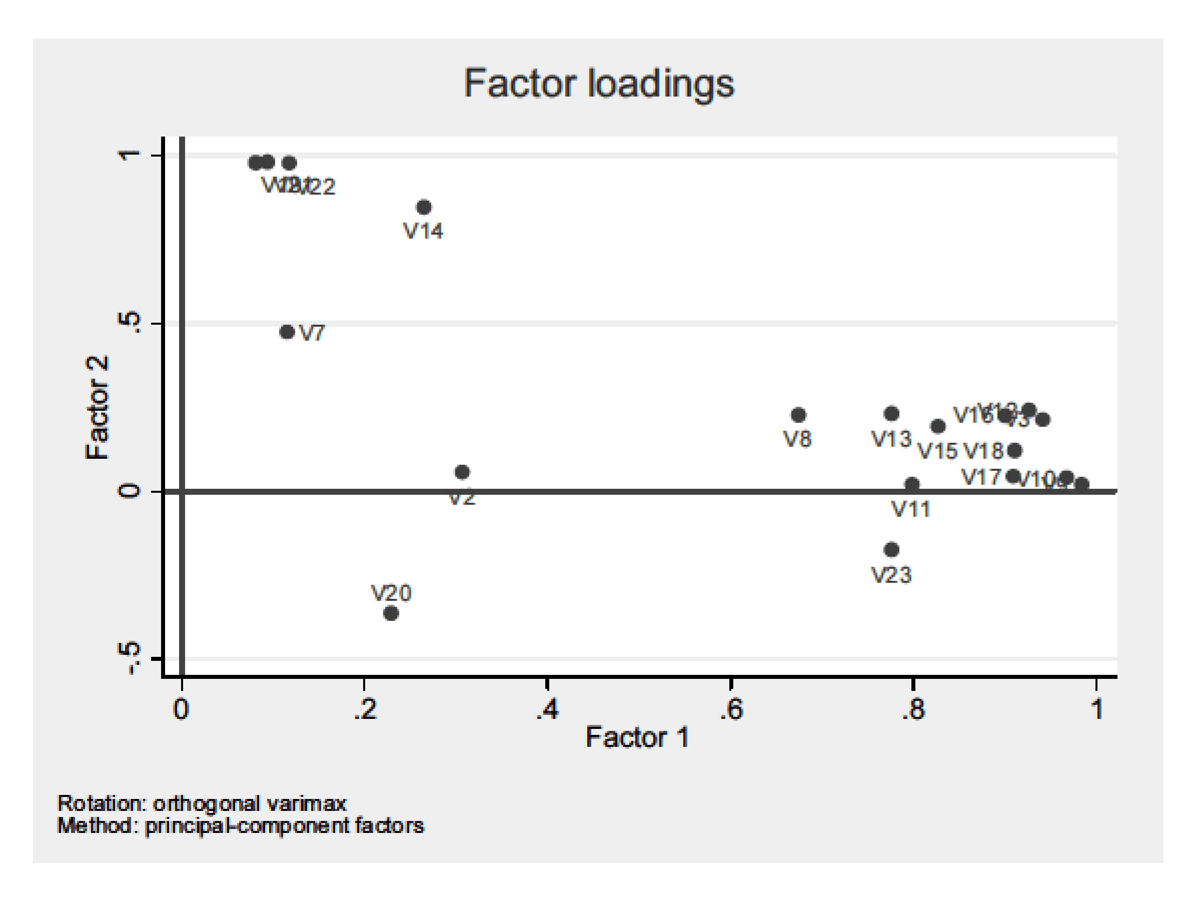
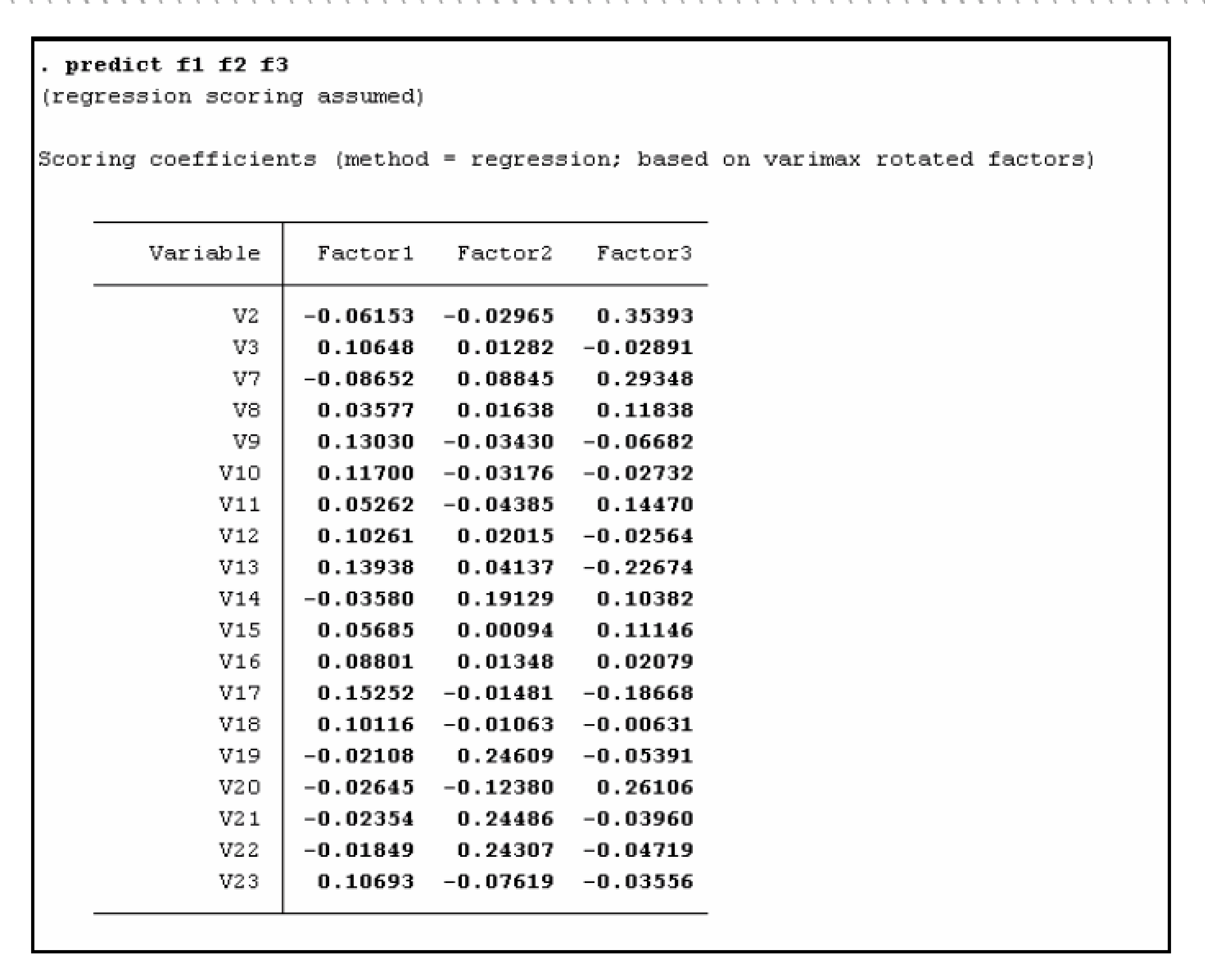


图 17.30　因子分析结果图 3

与前面的分析相同，我们发现 Factor1 主要解释的是 V3、V8、V9、V10~V13、V15~V18、V23 变量的信息，Factor2 主要解释的是 V14、V19、V21、V22 变量的信息。

图 17.31 展示的是因子分析后各个样本的因子得分情况。因子得分的概念是通过将每个变量标准化为平均数等于 0 和方差等于 1，然后以因子分析系数进行加权合计为每个因子构成的线性情况。以因子的方差贡献率为权数对因子进行加权求和，即可得到每个样本的因子综合得分。

根据图 17.31 展示的因子得分系数矩阵，可以写出各公因子的表达式。值得一提的是，在表达式中各个变量已经不是原始变量而是标准化变量。



```
. predict f1 f2 f3
(regression scoring assumed)

Scoring coefficients (method = regression; based on varimax rotated factors)
```

Variable	Factor1	Factor2	Factor3
V2	-0.06153	-0.02965	0.35393
V3	0.10648	0.01282	-0.02891
V7	-0.08652	0.08845	0.29348
V8	0.03577	0.01638	0.11838
V9	0.13030	-0.03430	-0.06682
V10	0.11700	-0.03176	-0.02732
V11	0.05262	-0.04385	0.14470
V12	0.10261	0.02015	-0.02564
V13	0.13938	0.04137	-0.22674
V14	-0.03580	0.19129	0.10382
V15	0.05685	0.00094	0.11146
V16	0.08801	0.01348	0.02079
V17	0.15252	-0.01481	-0.18668
V18	0.10116	-0.01063	-0.00631
V19	-0.02108	0.24609	-0.05391
V20	-0.02645	-0.12380	0.26106
V21	-0.02354	0.24486	-0.03960
V22	-0.01849	0.24307	-0.04719
V23	0.10693	-0.07619	-0.03556

图 17.31　因子分析结果图 4

表达式如下（只保留小数点后 3 位）：

F1=-0.062*年底总人口+0.106*地区生产总值-0.087*客运量+0.036*货运量
+0.130*地方财政预算内收入+0.117*地方财政预算内支出
+0.053*固定资产投资总额+0.103*城乡居民储蓄年末余额
+0.139*在岗职工平均工资-0.036*年末邮政局数+0.057*年末固定电话用户数
+0.088*社会商品零售总额+0.153*货物进出口总额
+0.101*年末实有公共汽车营运车辆数-0.021*影剧院数
-0.026*普通高等学校在校学生数-0.024*医院数-0.018*执业医师
+0.107*环境污染治理投资总额

F2=-0.030*年底总人口+0.013*地区生产总值+0.088*客运量+0.016*货运量
-0.034*地方财政预算内收入-0.032*地方财政预算内支出
-0.041*固定资产投资总额+0.020*城乡居民储蓄年末余额
+0.041*在岗职工平均工资+0.191*年末邮政局数
+0.001*年末固定电话用户数+0.013*社会商品零售总额
-0.015*货物进出口总额-0.011*年末实有公共汽车营运车辆数
+0.246*影剧院数-0.124*普通高等学校在校学生数+0.245*医院数
+0.243*执业医师-0.076*环境污染治理投资总额

F3=0.354*年底总人口-0.029*地区生产总值+0.293*客运量+0.118*货运量
-0.007*地方财政预算内收入-0.027*地方财政预算内支出
+0.145*固定资产投资总额-0.026*城乡居民储蓄年末余额
-0.227*在岗职工平均工资+0.104*年末邮政局数
+0.111*年末固定电话用户数+0.021*社会商品零售总额

-0.187*货物进出口总额+0.006*年末实有公共汽车营运车辆数

-0.054*影剧院数+0.261*普通高等学校在校学生数-0.040*医院数

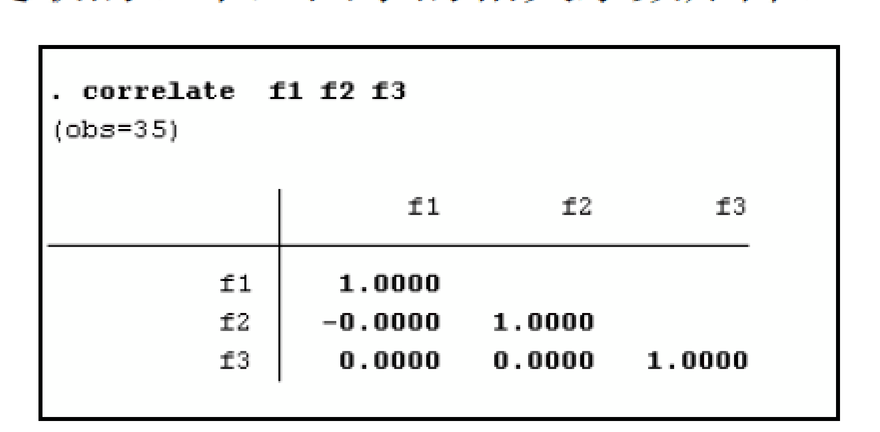
-0.047*执业医师-0.036*环境污染治理投资总额

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 17.32 所示的因子得分数据。

	V20	V21	V22	V23	yhat	e	f1	f2	f3
1	567875	686	54989	493300	9.41e+07	-575593.6	2.801096	-.2923421	.1634395
2	371136	411	26228	1217394	4.84e+07	2096997	.8953984	-.2125461	.1803138
3	316796	394	18773	0	2.19e+07	1726529	-.5646068	-.1963396	.4943747
4	298188	273	13646	234307	1.55e+07	-2902802	-.4487693	-.2129439	-.3796249
5	164998	146	6160	29348	1.08e+07	163034.9	-.5631225	-.132059	-.7737404
6	317450	314	19700	141618	3.12e+07	1055269	.0016073	-.2405628	.2252649
7	219982	236	14119	191069	3.18e+07	-507858.6	.0502157	-.1409505	-.0183759
8	331019	320	15670	0	1.93e+07	1607870	-.4243949	-.2299247	.0828417
9	384888	472	18000	341754	2.10e+07	3325525	-.3339622	-.2371567	.4639128
10	484873	534	48825	3661231	1.22e+08	-820.311	4.194826	-.7626205	-.145853
11	679924	167	15705	996994	3.42e+07	-1376059	.3043109	-.4343604	.4375622
12	366160	1162	20701	0	3.83e+07	2750722	.2266355	-.0071716	.159014
13	126094	266	15418	224191	3.34e+07	969433.4	.0644145	.0741792	-.1257256
14	295819	225	8539	43819	1.40e+07	-657705.8	-.4840672	-.2355267	-.2377106
15	233133	221	11601	74600	2.33e+07	-3513743	-.3993243	-.1310453	-.0475428
16	105545	53	5950	272202	1.42e+07	-286734.7	-.233629	-.1601152	-.9957065
17	481107	166	6363	61164	1.04e+07	3450905	-.5935123	-.3584638	-.0561947
18	570794	243	15800	0	2.48e+07	829734.2	-.3532399	-.3016038	.3898328
19	264917	251	15018	653008	3.70e+07	848750.6	.0991532	-.134754	.3176552
20	495719	307	16002	0	2.24e+07	2472236	-.3990123	-.28876	.5366887
21	778368	227	21541	72382	3.32e+07	-1755588	-.1216631	-.4432063	1.071549
22	454288	265	14683	158808	2.14e+07	509667.4	-.3364971	-.0424128	.3190583
23	265	29056	1484003	0	7.11e+07	414.1629	.4635315	5.635877	-.1069837
24	58910	101	18785	12534	6.83e+07	-264358.2	1.443625	.1050908	-1.664759
25	238375	201	11879	261235	1.12e+07	-494124.1	-.5788448	-.1729628	-.1787034
26	91152	75	4368	16831	5161705	-1224846	-.7429127	-.0083786	-.6926775
27	413655	1447	38739	619723	4.23e+07	-1053245	-.1783884	.4341447	4.066226
28	540626	595	28673	0	3.62e+07	-2929076	-.205659	-.0951683	1.547233
29	209499	268	10035	.	7264870	-327699.7	.	.	.

图 17.32 查看数据

图 17.33 展示的是系统提取的 3 个主因子的相关系数矩阵。



```
. correlate  f1 f2 f3
(obs=35)

             |       f1       f2       f3
-------------+---------------------------
          f1 |   1.0000
          f2 |  -0.0000   1.0000
          f3 |   0.0000   0.0000   1.0000
```

图 17.33 因子分析结果图 5

从图 17.33 中可以看出，提取的 3 个主因子之间几乎没有什么相关关系，这也说明了在前面对因子进行旋转的操作环节中采用最大方差正交旋转方式是明智的。值得说明的是图中有的相关系数是-0.0000 并非是不正确的，这是因为 Stata 14.0 只保留了 4 位小数所导致的，例如真实的数据有可能是-0.00001，那么结果显示的就是-0.0000。

图 17.34 展示的是每个样本在前两个主因子维度上的因子得分示意图。

从图 17.34 中可以看出，所有的样本被分到 4 个象限，可以比较直观地看出各个样本的因子得分分布情况。

图 17.35 展示的是本例因子分析的 KMO 检验结果。

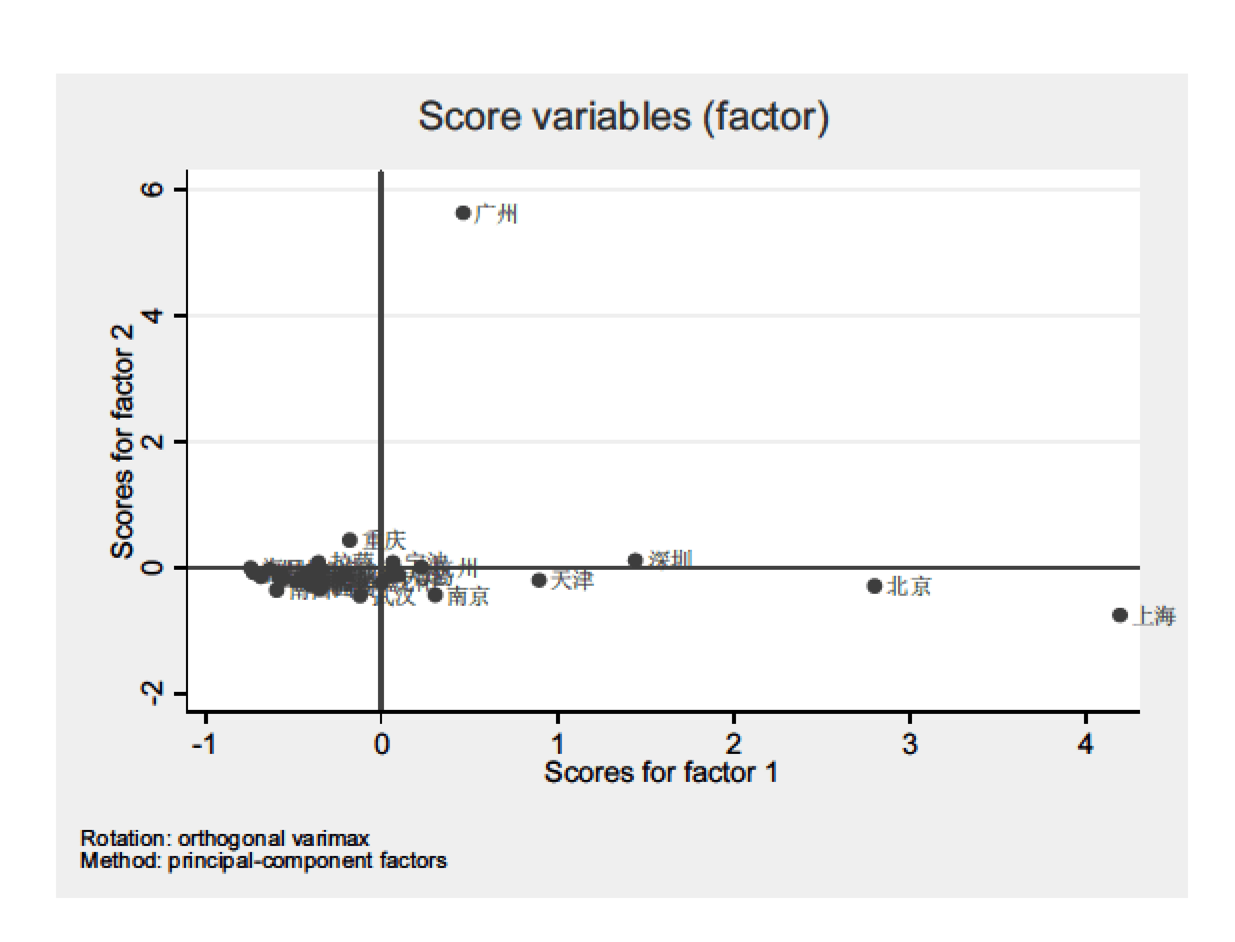


图 17.34　因子分析结果图 6

```
. estat kmo

Kaiser-Meyer-Olkin measure of sampling adequacy
```

Variable	kmo
V2	0.6519
V3	0.7686
V7	0.5885
V8	0.8298
V9	0.7998
V10	0.8218
V11	0.8748
V12	0.8614
V13	0.7619
V14	0.8424
V15	0.9412
V16	0.8360
V17	0.6747
V18	0.9112
V19	0.6267
V20	0.7363
V21	0.6802
V22	0.6503
V23	0.7828
Overall	0.7898

图 17.35　因子分析结果图 7

KMO 检验是为了查看数据是否适合进行因子分析，其取值范围是 0~1。其中，0.9~1 表示极好、0.8~0.9 表示可奖励的、0.7~0.8 表示还好、0.6~0.7 表示中等。本例中总体（Overall）KMO 的取值为 0.7898，表明因子分析的效果还是不错的。

图 17.36 展示的是本例因子分析所提取的各个因子的特征值碎石图。

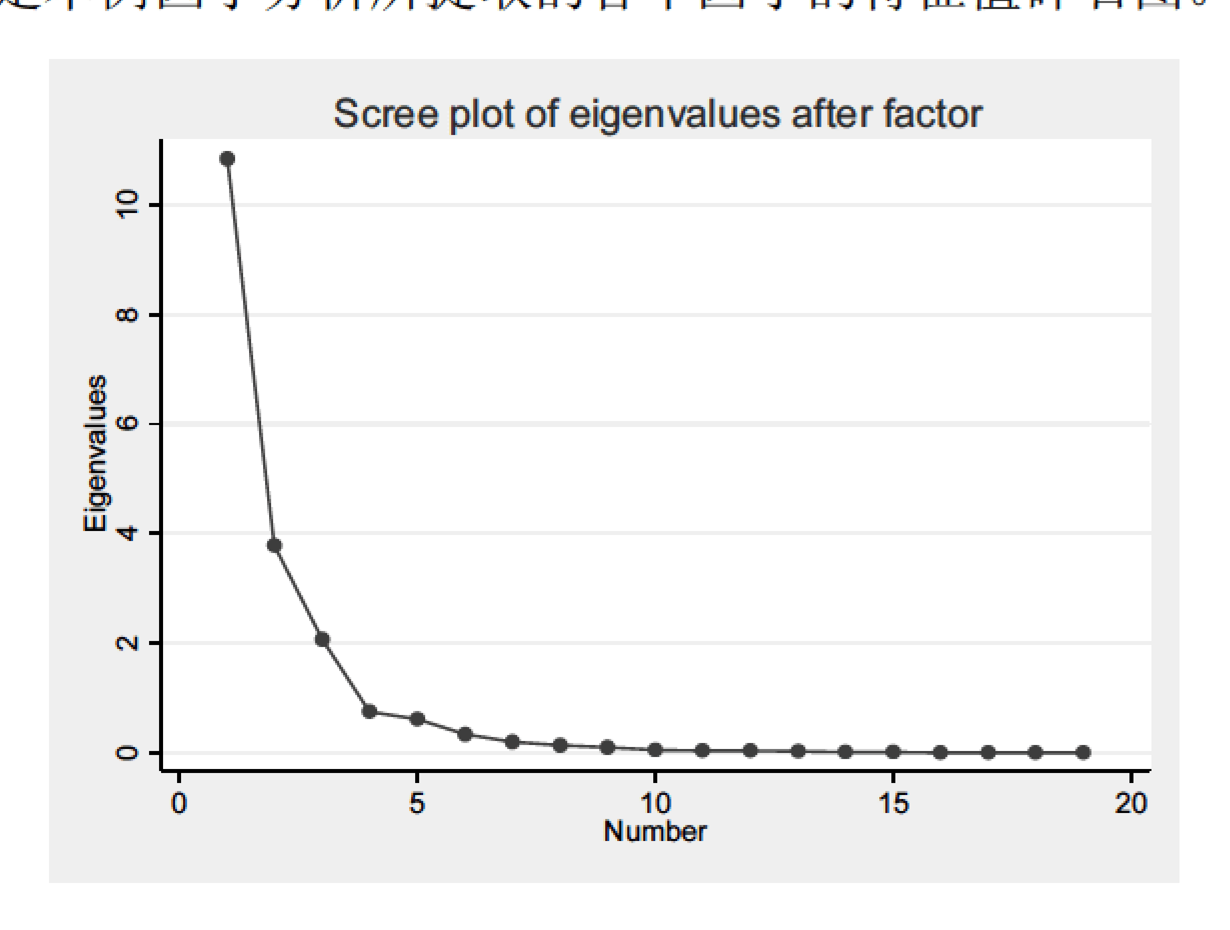


图 17.36　因子分析结果图 8

碎石图可以非常直观地观测出提取因子特征值的大小情况。图 17.36 的横轴表示的是系统提取因子的名称，并且已经按特征值大小进行降序排列，纵轴表示因子特征值的大小情况。从图 17.36 中可以轻松地看出本例中只有前 3 个因子的特征值是大于 1 的。

17.8 因子分析之后续分析

对于本部分分析，我们准备依照提取的公因子对各城市进行分类及排序。

操作步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中分别输入如下命令并按键盘上的回车键进行确认：generate f=0.4918*f1+0.2268*f2+0.1604*f3，本命令的含义是产生“综合得分”，这一变量将最终代表各个城市的综合经济实力，其中 f1、f2、f3 是在进行因子分析的时候对提取的公因子保存的变量，前面的系数是各个公因子的方差贡献率。

03 设置完毕后，等待输出结果。

在 Stata 14.0 主界面的结果窗口可以看到如图 17.37 和图 17.38 所示的分析结果。

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 17.37 所示的“综合得分”变量数据。

	V1	f1	f2	f3	f
1	北京	2.801096	-.2923421	.1634395	1.337492
2	天津	.8953984	-.2125461	.1803138	.4210738
3	石家	-.5646068	-.1963396	.4943747	-.2429058
4	太原	-.4487693	-.2129439	-.3796249	-.3298922
5	呼和	-.5631225	-.132059	-.7737404	-.4310026
6	沈阳	.0016073	-.2405628	.2252649	-.0176367
7	大连	.0502157	-.1409505	-.0183759	-.010219
8	长春	-.4243949	-.2299247	.0828417	-.2475765
9	哈尔	-.3339622	-.2371567	.4639128	-.1436182
10	上海	4.194826	-.7626205	-.145853	1.866658
11	南京	.3043109	-.4343604	.4375622	.1213321
12	杭州	.2266355	-.0071716	.159014	.1353386
13	宁波	.0644145	.0741792	-.1257256	.0283365
14	合肥	-.4840672	-.2355267	-.2377106	-.3296105
15	福州	-.3993243	-.1310453	-.0475428	-.2337346
16	厦门	-.273629	-.1601152	-.9957065	-.3109242
17	南昌	-.5935123	-.3584638	-.0561947	-.3822026
18	济南	-.3532399	-.3016038	.3898328	-.1795979
19	青岛	.0991532	-.134754	.3176552	.0691532
20	郑州	-.3990123	-.28876	.5366887	-.1756402
21	武汉	-.1216631	-.4432063	1.071549	.0115234
22	长沙	-.3364971	-.0424128	.3190583	-.1239316
23	广州	.4635315	5.635877	-.1069837	1.489021
24	深圳	1.443625	.1050908	-1.664759	.4667821
25	南宁	-.5788448	-.1729628	-.1787034	-.3525679
26	海口	-.7429127	-.0083786	-.6926775	-.4783702
27	重庆	-.1783884	.4341447	4.066226	.6629552
28	成都	-.205659	-.0951683	1.547233	.1254489
29	贵阳	.	.	.	.
30	昆明	-.4220155	-.1592225	-.2142413	-.2780232
31	拉萨	-.3586395	.0712314	-1.855197	-.4577973
32	西安	-.3484016	-.34375	.6554776	-.1441678
33	兰州	-.6845045	-.1517735	-.5732538	-.4630114
34	西宁	-.7266381	-.0932132	-.9776501	-.5353164
35	银川	-.6331403	-.0313378	-1.139654	-.5012863
36	乌鲁	-.4098681	-.0698505	-.9268498	-.3660819

图 17.37 因子分析之后续分析结果图 1

可以对数据进行排序操作，在主界面的“Command”文本框中输入操作命令：

```
sort f
```

并按键盘上的回车键进行确认，然后选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 17.38 所示的整理后的数据。

	V1	f1	f2	f3	f
1	西宁	-.7266381	-.0932132	-.9776501	-.5353164
2	银川	-.6331403	-.0313378	-1.139654	-.5012863
3	海口	-.7429127	-.0083786	-.6926775	-.4783702
4	兰州	-.6845045	-.1517735	-.5732538	-.4630114
5	拉萨	-.3586395	.0712314	-1.855197	-.4577973
6	呼和	-.5631225	-.132059	-.7737404	-.4310026
7	南昌	-.5935123	-.3584638	-.0561947	-.3822026
8	乌鲁	-.4098681	-.0698505	-.9268498	-.3660819
9	南宁	-.5788448	-.1729628	-.1787034	-.3525679
10	太原	-.4487693	-.2129439	-.3796249	-.3298922
11	合肥	-.4840672	-.2355267	-.2377106	-.3296105
12	厦门	-.233629	-.1601152	-.9957065	-.3109242
13	昆明	-.4220155	-.1592225	-.2142413	-.2780232
14	长春	-.4243949	-.2299247	.0828417	-.2475765
15	石家	-.5646068	-.1963396	.4943747	-.2429058
16	福州	-.3993243	-.1310453	-.0475428	-.2337346
17	济南	-.3532399	-.3016038	.3898328	-.1795979
18	郑州	-.3990123	-.28876	.5366887	-.1756402
19	西安	-.3484016	-.34375	.6554776	-.1441678
20	哈尔	-.3339622	-.2371567	.4639128	-.1436182
21	长沙	-.3364971	-.0424128	.3190583	-.1239316
22	沈阳	.0016073	-.2405628	.2252649	-.0176367
23	大连	.0502157	-.1409505	-.0183759	-.010219
24	武汉	-.1216631	-.4432063	1.071549	.0115234
25	宁波	.0644145	.0741792	-.1257256	.0283365
26	青岛	.0991532	-.134754	.3176552	.0691532
27	南京	.3043109	-.4343604	.4375622	.1213321
28	成都	-.205659	-.0951683	1.547233	.1254489
29	杭州	.2266355	-.0071716	.159014	.1353386
30	天津	.8953984	-.2125461	.1803138	.4210738
31	深圳	1.443625	.1050908	-1.664759	.4667821
32	重庆	-.1783884	.4341447	4.066226	.6629552
33	北京	2.801096	-.2923421	.1634395	1.337492
34	广州	.4635315	5.635877	-.1069837	1.489021
35	上海	4.194826	-.7626205	-.145853	1.866658
36	贵阳	.	.	.	.

图 17.38　因子分析之后续分析结果图 2

观察综合得分列可以发现：除贵阳因数据缺失未参加排名外，上海“一骑绝尘，一枝独秀”，是中国综合经济实力最强的城市；北京、广州两个城市综合得分紧随其后，综合经济实力也是很强的，与上海构成前三甲；武汉、宁波、南京、青岛、成都、深圳、天津、重庆、杭州等城市的综合得分在 0~1 之间，综合经济实力较强；大连、沈阳、长沙、哈尔滨、西安、济南、厦门、郑州、福州、长春、昆明、乌鲁木齐、石家庄、太原、拉萨、合肥、南宁、呼和浩特、南昌、银川、兰州、海口、西宁等城市的综合得分均为负值，综合经济实力相对较弱，其中最弱的是西宁，得分为-0.54。所有城市的综合经济实力排名依次为：上海、广州、北京、重庆、深圳、天津、杭州、成都、南京、青岛、宁波、武汉、大连、沈阳、长沙、哈尔滨、西安、郑州、济南、福州、石家庄、长春、昆明、厦门、合肥、太原、南宁、乌鲁木齐、南昌、呼和浩特、拉萨、兰州、海口、银川、西宁。

17.9　研究结论

- 简单相关分析表明：构成“地区生产总值”的 3 个组成部分只有“第二产业增加值”与“第三产业增加值”之间具有很强的相关性，并且在 0.01 的显著性水平上显著，其他的变量之间相关性很不显著。
- 简单相关分析表明：“客运量”与“货运量”之间虽然相关系数不是很大，但是这种相关性却很强，在 0.01 的显著性水平上显著。
- 简单相关分析表明：“地方财政预算内收入”和“地方财政预算内支出”相关系数不

是很大，而且这种相关性很强，在 0.01 的显著性水平上显著。

- 简单相关分析表明：年底总人口与地区生产总值为正相关但相关系数不大；年底总人口与环境污染治理投资总额之间也为正相关而且相关系数也不大；地区生产总值与环境污染治理投资总额之间也为正相关，而且相关系数较大。只有地区生产总值与环境污染治理投资总额之间的相关关系非常显著（在 0.01 的水平上显著）。
- 经过多重线性回归分析，可以发现我国城市的地区生产总值与社会商品零售总额、货物进出口总额、货运量、固定资产投资总额、年末邮政局数、影剧院数、医院数、年末实有公共汽车营运车辆数有显著关系，与其他变量之间的关系并不显著。其中固定资产投资总额、社会商品零售总额、货物进出口总额、医院数、货运量对地区生产总值起正向作用，尤其是医院数和货运量，每增加一个单位，地区生产总值就分别增加 4141.523 个单位和 196.4853 个单位，而年末邮政局数、影剧院数、年末实有公共汽车营运车辆数对地区生产总值起反向作用。
- 可以用 3 个公因子来概括所有描述我国城市综合经济实力的指标：第 1 个因子用来反映地区生产总值、货运量、地方财政预算内收入、地方财政预算内支出、固定资产投资总额、城乡居民储蓄年末余额、在岗职工平均工资、年末固定电话用户数、社会商品零售总额、货物进出口总额、年末实有公共汽车营运车辆数、环境污染治理投资总额等变量的信息；第 2 个因子用来反映年末邮政局数、影剧院数、医院数、执业医师等变量的信息；第 3 个因子用来反映年底总人口、客运量、普通高等学校在校学生数等变量的信息。
- 因子分析之后续分析表明，所有城市的综合经济实力排名依次为：上海、广州、北京、重庆、深圳、天津、杭州、成都、南京、青岛、宁波、武汉、大连、沈阳、长沙、哈尔滨、西安、郑州、济南、福州、石家庄、长春、昆明、厦门、合肥、太原、南宁、乌鲁木齐、南昌、呼和浩特、拉萨、兰州、海口、银川、西宁。

经过以上研究，我们可以从一种宏观的视野上对我国的城市综合经济实力有一个比较全面的了解，这对于以后我国城市的发展有着重要的借鉴和指导意义。例如，根据回归分析部分的结论，为提高地区生产总值，我国各城市必须要积极扩大货运量，“要想富，先修路”这句话是非常有道理的。再如，因子分析之后续分析表明，排名在前的大多是东部城市，在后的基本上都是中西部城市，由于城市经济往往代表着一个地区的先进生产力，所以为使我国经济均衡发展，加强中西部建设是非常有必要的。

17.10　本章习题

使用《中国统计年鉴 2007》上的《中国 2006 年省会城市和计划单列市主要经济指标统计（包括市辖县）》数据（数据已整理至 Stata 中），进行以下分析。

（1）相关分析

- 对“地区生产总值”和“工业增加值”进行简单相关分析。
- 对“客运量”和“货运量”进行简单相关分析。

- 对“地方财政预算内收入”和“地方财政预算内支出”进行简单相关分析。
- 对“年底总人口”“地区生产总值”“环境污染治理投资总额”这 3 个变量进行简单相关分析。

（2）回归分析

以“地区生产总值”为因变量，以“年底总人口”“客运量”“货运量”“地方财政预算内收入”“地方财政预算内支出”“固定资产投资总额”“城乡居民储蓄年末余额”“在岗职工平均工资”“年末邮政局数”“年末固定电话用户数”“社会商品零售总额”“货物进出口总额”“年末实有公共汽车营运车辆数”“影剧院数”“普通高等学校在校学生数”“医院数”“执业医师”“环境污染治理投资总额”等为自变量，进行多重线性回归。

（3）因子分析

对构成城市综合经济实力的各个变量（“年底总人口”“地区生产总值”“客运量”“货运量”“地方财政预算内收入”“地方财政预算内支出”“固定资产投资总额”“城乡居民储蓄年末余额”“在岗职工平均工资”“年末邮政局数”“年末固定电话用户数”“社会商品零售总额”“货物进出口总额”“年末实有公共汽车营运车辆数”“影剧院数”“普通高等学校在校学生数”“医院数”“执业医师”“环境污染治理投资总额”）提取公因子。

（4）因子分析之后续分析

依照提取的公因子对各城市进行分类及排序。

第 18 章　Stata 在旅游业中的应用

旅游业作为第三产业的重要组成部分，是世界上发展最快的新兴产业之一。它一方面能够满足人们日益增长的物质和文化的需要，另一方面又直接或者间接地促进国民经济有关部门的发展。随着社会的发展，旅游业在国民经济中的地位越来越重要，也越来越引起政府官员和社会学者的重点关注。本章就来介绍一下 Stata 在对旅游业研究中的应用。

18.1　研究背景及目的

背景一：进入 21 世纪以来，中国旅游业快速发展，旅游人数迅速增加。

根据《中国投资年鉴 2007》提供的数据（表 18.1）可以发现，除 2003 年稍有下降外，无论是国内旅游人数还是入境旅游人数都呈现出不断递增的趋势。

表 18.1　国内旅游人数和入境旅游人数统计（2001—2006 年）

年份	2001年	2002年	2003年	2004年	2005年	2006年
国内旅游人数/亿人次	7.84	8.78	8.70	11.02	12.12	13.94
入境旅游人数/万人次	8 901.29	9 790.83	9 166.21	10 903.82	12 029.23	12 494.21

背景二：伴随着旅游人数的不断增加，我国的旅行社个数和星级饭店数增长迅速。

根据《中国投资年鉴 2007》提供的数据（表 18.2）可以发现，从 2001 年到 2006 年，旅行社个数和星级饭店个数不断递增。

表 18.2　旅行社个数和星级饭店个数统计（2001—2006 年）

年份	2001年	2002年	2003年	2004年	2005年	2006年
旅行社个数	10 532	11 552	13 361	14 927	16 245	18 475
星级饭店个数	7 358	8 880	9 751	10 888	11 828	12 494

背景三：伴随着旅游人数、旅行社个数的增加，旅游收入不断增长，而且速度很快。

根据《中国投资年鉴 2007》提供的数据（表 18.3）可以发现，除 2003 年稍有下降外，无论是国际旅游收入还是国内旅游收入都呈现出不断递增的趋势。

表 18.3　旅游收入统计（2001—2006 年）

年份	2001年	2002年	2003年	2004年	2005年	2006年
国际旅游收入/亿美元	187.92	203.85	184.06	257.39	292.96	339.49
国内旅游收入/亿元	3 522.36	3 878.36	3 442.27	4 710.71	5 285.86	6 229.74

一般来说，旅游消费的地域差异不但是地区经济发展不平衡的集中表现和缩影，而且反映着地区间文化和人民消费特点的差异，所以从这两个角度来说，按照不同的分类指标对我国

各地区居民的人均旅游消费支出进行分解分析研究，并且从量上明确我国居民旅游消费性支出的区域差异，具有非常重大的意义。

18.2 研究方法

本例采用的数据有《中国2007年城镇居民国内旅游出游人均花费情况统计（按城市、性别和年龄分组）》《中国2007年城镇居民国内旅游出游人均花费情况统计（按城市和家庭月平均收入分组）》《中国2007年城镇居民国内旅游出游人均花费情况统计（按城市和旅游目的分组）》《中国2007年城镇居民国内旅游出游人均花费情况统计（按城市和文化程度分组）》《中国2007年城镇居民国内旅游出游人均花费情况统计（按城市和职业分组）》《中国2007年国家级风景名胜区统计》等，这些数据都摘自《中国国内旅游抽样调查资料2008》。

因为我们研究的主要目的是找出各地区的相应指标或数据之间存在的相似性或相异性，所以主要采用聚类分析方法对相关数据展开分析。聚类分析是采用定量数学方法，根据样品或指标的数值特征，对样品进行分类来推断各样品之间亲疏关系的一种分析方法。

基本思路是：一方面，针对中国2007年城镇居民国内旅游出游人均花费情况的各种不同分类分别使用聚类分析对各地区进行聚类；另一方面，使用聚类分析方法对中国2007年部分国家级风景名胜区进行聚类。

18.3 数据分析与报告

18.3.1 各城市国内旅游出游人均花费按性别和年龄进行的聚类分析

	下载资源:\video\chap18\…
	下载资源:\sample\chap18\案例18.1.dta

表18.4是2007年中国22个城市城镇居民国内旅游出游人均花费按性别和年龄进行分类的数据。

表18.4 中国2007年城镇居民国内旅游出游人均花费情况统计（按性别和年龄分组）（单位：元/人）

城市	性别		年龄				
	男	女	65岁及以上	45~65岁	25~44岁	15~24岁	0~14岁
北京	1 051.0	1 032.8	1 011.5	958.2	1 290.8	1 052.1	603.8
天津	895.8	767.8	714.9	918.9	895.1	4 86.8	598.7
石家庄	925.7	715.1	1 184.7	1 050.0	637.1	1 254.2	336.1
太原	1 818.9	1 402.5	1 965.7	1 938.6	1 290.6	1 100.5	616.0
呼和浩特	2 306.5	1 880.9	2 574.5	2 568.9	1 679.6	973.5	1 096.7
沈阳	388.3	469.8	505.2	465.2	405.8	375.4	272.4
大连	328.8	344.5	437.7	358.7	339.2	302.1	183.8

（续表）

城市	性别		年龄				
	男	女	65岁及以上	45~65岁	25~44岁	15~24岁	0~14岁
长春	2 221.6	2 956.7	2 387.5	3 187.6	2 218.5	2 600.0	1 864.5
哈尔滨	2 477.2	1 459.4	1 289.8	2 807.1	1 423.7	983.1	372.8
上海	1 103.6	706.4	485.5	910.0	1 032.6	640.4	670.8
南京	2 441.1	2 185.2	1 641.1	2605.0	2 327.6	2 197.9	1 560.2
无锡	1 070.3	1 059.8	459.0	855.6	1 492.6	950.0	469.3
苏州	762.4	647.4	544.3	924.1	616.7	180.5	332.2
杭州	1 000.1	832.5	683.1	1 041.6	769.2	1 622.3	393.0
青岛	1 397.1	1 016.7	1 599.1	925.9	1 384.3	1 549.2	419.1
郑州	921.3	825.3	1 408.0	946.1	865.7	438.5	628.3
武汉	988.5	784.9	620.9	900.4	996.4	733.1	431.5
长沙	1 191.4	1 445.2	904.6	1 559.2	1 382.7	1 446.0	711.3
广州	777.5	846.4	473.9	830.1	977.0	690.2	442.0
深圳	2 923.3	2 613.5	983.5	2 996.7	2 947.8	1 926.8	1 064.5
银川	1 473.1	1 441.4	382.1	1 446.4	1 648.4	1 124.5	1 210.2
乌鲁木齐	1 200.9	1 166.0	2 744.9	1 454.4	1 182.4	834.8	584.2

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有 8 个变量，分别为“城市”“男”“女”“65 岁及以上”“45~65 岁”“25~44 岁”“15~24 岁”“0~14 岁”。我们将这 8 个变量分别定义为 V1~V8，然后录入相关数据。录入完成后数据如图 18.1 所示。

	V1	V2	V3	V4	V5	V6	V7	V8
1	北京	1051	1032.8	1011.5	958.2	1290.8	1052.1	603.8
2	天津	895.8	767.8	714.9	918.9	895.1	486.8	598.7
3	石家庄	925.7	715.1	1174.7	1050	637.1	1254.2	336.1
4	太原	1717.9	1402.5	1965.7	1938.6	1290.6	1100.5	616
5	呼和浩特	2306.5	1880.9	2574.5	2568.9	1679.6	973.5	1096.7
6	沈阳	388.3	469.8	505.2	465.2	405.8	375.4	272.4
7	大连	328.8	344.5	437.7	358.7	339.2	302.1	173.8
8	长春	2221.6	2956.7	2387.5	3187.6	2217.5	2600	1864.5
9	哈尔滨	2477.2	1459.4	1289.8	2807.1	1423.7	983.1	372.8
10	上海	1103.6	706.4	485.5	910	1032.6	640.4	670.8
11	南京	2441.1	2175.2	1641.1	2605	2327.6	2197.9	1560.2
12	无锡	1070.3	1059.8	459	855.6	1492.6	950	469.3
13	苏州	762.4	647.4	544.3	924.1	616.7	170.5	332.2
14	杭州	1000.1	832.5	683.1	1041.6	769.2	1622.3	393
15	青岛	1397.1	1016.7	1599.1	925.9	1384.3	1549.2	419.1
16	郑州	921.3	825.3	1408	946.1	865.7	438.5	628.3
17	武汉	988.5	784.9	620.9	900.4	996.4	733.1	431.5
18	长沙	1191.4	1445.2	904.6	1559.2	1382.7	1446	711.3
19	广州	777.5	846.4	473.9	830.1	977	690.2	442
20	深圳	2923.3	2613.5	983.5	2996.7	2947.8	1926.8	1064.5
21	银川	1473.1	1441.4	382.1	1446.4	1648.4	1124.5	1210.2
22	乌鲁木齐	1200.9	1166	2744.9	1454.4	1182.4	834.8	584.2

图 18.1　案例 18.1 数据

聚类分析的分析步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中分别输入如下命令并按键盘上的回车键进行确认。

- egen zv2=std(V2)：本命令旨在对 V2 变量数据进行标准化处理，标准化处理方式是使变量的平均数为 0 而且标准差为 1。
- egen zv3=std(V3)：本命令旨在对 V3 变量数据进行标准化处理，标准化处理方式是使变量的平均数为 0 而且标准差为 1。
- egen zv4=std(V4)：本命令旨在对 V4 变量数据进行标准化处理，标准化处理方式是使变量的平均数为 0 而且标准差为 1。
- egen zv5=std(V5)：本命令旨在对 V5 变量数据进行标准化处理，标准化处理方式是使

变量的平均数为 0 而且标准差为 1。

- egen zv6=std(V6)：本命令旨在对 V6 变量数据进行标准化处理，标准化处理方式是使变量的平均数为 0 而且标准差为 1。
- egen zv7=std(V7)：本命令旨在对 V7 变量数据进行标准化处理，标准化处理方式是使变量的平均数为 0 而且标准差为 1。
- egen zv8=std(V8)：本命令旨在对 V8 变量数据进行标准化处理，标准化处理方式是使变量的平均数为 0 而且标准差为 1。
- cluster kmeans zv2 zv3 zv4 zv5 zv6 zv7 zv8,k(2)：本命令旨在对 V2~V8 的标准化变量进行“K 个平均数的聚类分析”，设定的聚类数是 2。
- cluster kmeans zv2 zv3 zv4 zv5 zv6 zv7 zv8,k(3)：本命令旨在对 V2~V8 的标准化变量进行“K 个平均数的聚类分析”，设定的聚类数是 3。
- cluster kmeans zv2 zv3 zv4 zv5 zv6 zv7 zv8,k(4)：本命令旨在对 V2~V8 的标准化变量进行“K 个平均数的聚类分析”，设定的聚类数是 4。

03 设置完毕后，按键盘上的回车键，等待输出结果。

在 Stata 14.0 主界面的结果窗口可以看到如图 18.2~图 18.11 所示的分析结果。

1．数据标准化处理

在分析过程中前 7 条 Stata 命令旨在对数据进行标准化处理，选择的标准化处理方式是使变量的平均数为 0 而且标准差为 1。之所以这样做是因为进行聚类分析的变量都是以不可比的单位进行的测度，它们具有极为不同的方差，对数据进行标准化处理可以避免使结果受到具有最大方差变量的影响。在输入前 7 条 Stata 命令并且分别按键盘上的回车键进行确认后，选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 18.2 所示的变换后的数据。

	V6	V7	V8	zv2	zv3	zv4	zv5	zv6	zv7	zv8
1	1290.8	1052.1	603.8	-.4140195	-.2598782	-.1693536	-.5639904	.0425931	-.0222234	-.1658237
2	895.1	486.8	598.7	-.6334792	-.6515161	-.5728834	-.6101308	-.5807921	-.9263219	-.1776909
3	637.1	1254.2	336.1	-.5911993	-.7294003	.0526829	-.456212	-.9872448	.3010002	-.7887362
4	1290.6	1100.5	616	.529007	.2864935	1.128853	.5870553	.0422779	.055184	-.1374355
5	1679.6	973.5	1096.7	1.361314	.9935107	1.957137	1.327063	.6551079	-.1479303	.9811078
6	405.8	375.4	272.4	-1.351107	-1.091924	-.8581841	-1.142801	-1.351635	-1.104487	-.9369602
7	339.2	302.1	173.8	-1.435243	-1.277102	-.9500191	-1.267838	-1.456556	-1.221717	-1.166393
8	2217.5	2600	1864.5	1.241261	2.583412	1.70272	2.053453	1.502515	2.453372	2.767706
9	1423.7	983.1	372.8	1.602691	.3705848	.2092787	1.606724	.2519634	-.1325768	-.7033389
10	1032.6	640.4	670.8	-.3396407	-.7422578	-.8849863	-.62058	-.3641748	-.6806655	-.0099211
11	2327.6	2197.9	1560.2	1.551644	1.42845	.6872288	1.369447	1.675966	1.810283	2.059628
12	1492.6	950	469.3	-.3867283	-.2199755	-.9210401	-.6844487	.3605084	-.1855144	-.4787925
13	616.7	170.5	332.2	-.8221127	-.8294526	-.8049878	-.6040258	-1.019383	-1.432188	-.7978112
14	769.2	1622.3	393	-.4859944	-.5558974	-.6161479	-.4660741	-.7791347	.8897119	-.6563354
15	1384.3	1549.2	419.1	.0753814	-.2836721	.630087	-.6019124	.1898928	.7728011	-.5956031
16	865.7	438.5	628.3	-.5974211	-.5665381	.3700919	-.5781965	-.6271088	-1.003569	-.1088146
17	996.4	733.1	431.5	-.5023973	-.6262443	-.7007721	-.6318509	-.4212042	-.5324081	-.5667495
18	1382.7	1446	711.3	-.2154876	.3495988	-.3147931	.1416179	.1873721	.6077508	.0843186
19	977	690.2	442	-.8007607	-.5353548	-.9007683	-.7143872	-.4517669	-.6010191	-.542317
20	2947.8	1926.8	1064.5	2.233497	2.076205	-.2074481	1.829325	2.653028	1.376706	.9061816
21	1648.4	1124.5	1210.2	.1828488	.343983	-1.025664	.0091844	.6059555	.0935678	1.245212
22	1182.4	834.8	584.2	-.2020541	-.0630248	2.188969	.0185768	-.1281801	-.3697567	-.211431

图 18.2　按性别和年龄进行的聚类分析结果图 1

2．K 个平均数的聚类分析

图 18.3 展示的是设定聚类数为 2，然后使用“K 个平均数的聚类分析”方法进行分析的结果。在输入第 8 条 Stata 命令并且分别按键盘上的回车键进行确认后，可以看到系统产生了一

个新的聚类变量。

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 18.4 所示的聚类数据。

```
. cluster kmeans zv2 zv3 zv4 zv5 zv6 zv7 zv8,k(2)
cluster name: _clus_1
```

图 18.3　按性别和年龄进行的聚类分析结果图 2

	V1	zv3	zv4	zv5	zv6	zv7	zv8	_clus_2
1	北京	-.2598782	-.1693536	-.5639904	.0425931	-.0222234	-.1658237	1
2	天津	-.6515161	-.5728834	-.6101308	-.5807921	-.9263219	-.1776909	1
3	石家庄	-.7294003	.0526829	-.456212	-.9872448	.3010002	-.7887362	1
4	太原	.2864935	1.128853	.5870553	.0422779	.055184	-.1374355	1
5	呼和浩特	.9935107	1.957137	1.327063	.6551079	-.1479303	.9811078	2
6	沈阳	-1.091924	-.8581841	-1.142801	-1.351635	-1.104487	-.9369602	1
7	大连	-1.277102	-.9500191	-1.267838	-1.456556	-1.221717	-1.166393	1
8	长春	2.583412	1.70272	2.053453	1.502515	2.453372	2.767706	2
9	哈尔滨	.3705848	.2092787	1.606724	.2519634	-.1325768	-.7033389	2
10	上海	-.7422578	-.8849863	-.62058	-.3641748	-.6806655	-.0099211	1
11	南京	1.42845	.6872288	1.369447	1.675966	1.810283	2.059628	2
12	无锡	-.2199755	-.9210401	-.6844487	.3605084	-.1855144	-.4787925	1
13	苏州	-.8294526	-.8049878	-.6040258	-1.019383	-1.432188	-.7978112	1
14	杭州	-.5558974	-.6161479	-.4660741	-.7791347	.8897119	-.6563354	1
15	青岛	-.2836721	.630087	-.6019124	.1898928	.7728011	-.5956031	1
16	郑州	-.5665381	.3700919	-.5781965	-.6271088	-1.003569	-.1088146	1
17	武汉	-.6262443	-.7007721	-.6318509	-.4212042	-.5324081	-.5667495	1
18	长沙	.3495988	-.3147931	.1416179	.1873721	.6077508	.0843186	1
19	广州	-.5353548	-.9007683	-.7143872	-.4517669	-.6010191	-.542317	1
20	深圳	2.076205	-.2074481	1.829325	2.653028	1.376706	.9061816	2
21	银川	.343983	-1.025664	.0091844	.6059555	.0935678	1.245212	1
22	乌鲁木齐	-.0630248	2.188969	.0185768	-.1281801	-.3697567	-.211431	1

图 18.4　按性别和年龄进行的聚类分析结果图 3

在图 18.4 中，可以看到所有的观测样本被分为两类，其中呼和浩特、哈尔滨、长春、南京、深圳被分到第 2 类，其他的省市被分到第 1 类。

为观测两类样本的特征，可以对数据进行排序操作，在主界面的“Command”文本框中输入操作命令：

```
sort _clus_1
```

并按键盘上的回车键进行确认，然后选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 18.5 所示的整理后的数据。

	V1	zv3	zv4	zv5	zv6	zv7	zv8	_clus_1
1	青岛	-.2836721	.630087	-.6019124	.1898928	.7728011	-.5956031	1
2	杭州	-.5558974	-.6161479	-.4660741	-.7791347	.8897119	-.6563354	1
3	沈阳	-1.091924	-.8581841	-1.142801	-1.351635	-1.104487	-.9369602	1
4	北京	-.2598782	-.1693536	-.5639904	.0425931	-.0222234	-.1658237	1
5	郑州	-.5665381	.3700919	-.5781965	-.6271088	-1.003569	-.1088146	1
6	长沙	.3495988	-.3147931	.1416179	.1873721	.6077508	.0843186	1
7	上海	-.7422578	-.8849863	-.62058	-.3641748	-.6806655	-.0099211	1
8	银川	.343983	-1.025664	.0091844	.6059555	.0935678	1.245212	1
9	武汉	-.6262443	-.7007721	-.6318509	-.4212042	-.5324081	-.5667495	1
10	乌鲁木齐	-.0630248	2.188969	.0185768	-.1281801	-.3697567	-.211431	1
11	苏州	-.8294526	-.8049878	-.6040258	-1.019383	-1.432188	-.7978112	1
12	广州	-.5353548	-.9007683	-.7143872	-.4517669	-.6010191	-.542317	1
13	石家庄	-.7294003	.0526829	-.456212	-.9872448	.3010002	-.7887362	1
14	大连	-1.277102	-.9500191	-1.267838	-1.456556	-1.221717	-1.166393	1
15	无锡	-.2199755	-.9210401	-.6844487	.3605084	-.1855144	-.4787925	1
16	太原	.2864935	1.128853	.5870553	.0422779	.055184	-.1374355	1
17	天津	-.6515161	-.5728834	-.6101308	-.5807921	-.9263219	-.1776909	1
18	哈尔滨	.3705848	.2092787	1.606724	.2519634	-.1325768	-.7033389	1
19	长春	2.583412	1.70272	2.053453	1.502515	2.453372	2.767706	2
20	南京	1.42845	.6872288	1.369447	1.675966	1.810283	2.059628	2
21	呼和浩特	.9935107	1.957137	1.327063	.6551079	-.1479303	.9811078	2
22	深圳	2.076205	-.2074481	1.829325	2.653028	1.376706	.9061816	2

图 18.5　按性别和年龄进行的聚类分析结果图 4

可以看到第 2 类所代表的人均旅游消费支出特点是无论男女老少花费支出总体上相对较高，第 1 类所代表的人均旅游消费支出特点是无论男女老少各年龄段花费支出总体上相对较低。

图 18.6 展示的是设定聚类数为 3，然后使用“K 个平均数的聚类分析”方法进行分析的结果。在输入第 9 条 Stata 命令并且分别按键盘上的回车键进行确认后，可以看到系统产生了一个新的变量：聚类变量_clus_2（cluster name: _clus_2）。

```
. cluster kmeans zv2 zv3 zv4 zv5 zv6 zv7 zv8,k(3)
cluster name: _clus_2
```

图 18.6　按性别和年龄进行的聚类分析结果图 5

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 18.7 所示的_clus_2 数据。

从图 18.7 中可以看到，所有的观测样本被分为 3 类，其中长春、南京、呼和浩特、深圳属于第 2 类，青岛、长沙、银川、乌鲁木齐、太原、哈尔滨属于第 3 类，其他城市属于第 1 类。

	v1	zv3	zv4	zv5	zv6	zv7	zv8	_clus_1	_clus_2
1	青岛	-.2836721	.630087	-.6019124	.1898928	.7728011	-.5956031	1	3
2	杭州	-.5558974	-.6161479	-.4660741	-.7791347	.8897119	-.6563354	1	1
3	沈阳	-1.091924	-.8581841	-1.142801	-1.351635	-1.104487	-.9369602	1	1
4	北京	-.2598782	-.1693536	-.5639904	.0425931	-.0222234	-.1658237	1	1
5	郑州	-.5665381	.3700919	-.5781965	-.6271088	-1.003569	-.1088146	1	1
6	长沙	.3495988	-.3147931	.1416179	.1873721	.6077508	.0843186	1	3
7	上海	-.7422578	-.8849863	-.62058	-.3641748	-.6806655	-.0099211	1	1
8	银川	.343983	-1.025664	.0091844	.6059555	.0935678	1.245212	1	3
9	武汉	-.6262443	-.7007721	-.6318509	-.4212042	-.5324081	-.5667495	1	1
10	乌鲁木齐	-.0630248	2.188969	.0185768	-.1281801	-.3697567	-.211431	1	3
11	苏州	-.8294526	-.8049878	-.6040258	-1.019383	-1.432188	-.7978112	1	1
12	广州	-.5353548	-.9007683	-.7143872	-.4517669	-.6010191	-.542317	1	1
13	石家庄	-.7294003	.0526829	-.456212	-.9872448	.3010002	-.7887362	1	1
14	大连	-1.277102	-.9500191	-1.267838	-1.456556	-1.221717	-1.166393	1	1
15	无锡	-.2199755	-.9210401	-.6844487	.3605084	-.1855144	-.4787925	1	1
16	太原	.2864935	1.128853	.5870553	.0422779	.055184	-.1374355	1	3
17	天津	-.6515161	-.5728834	-.6101308	-.5807921	-.9263219	-.1776909	1	1
18	哈尔滨	.3705848	.2092787	1.606724	.2519634	-.1325768	-.7033389	1	3
19	长春	2.583412	1.70272	2.053453	1.502515	2.453372	2.767706	2	2
20	南京	1.42845	.6872288	1.369447	1.675966	1.810283	2.059628	2	2
21	呼和浩特	.9935107	1.957137	1.327063	.6551079	-.1479303	.9811078	2	2
22	深圳	2.076205	-.2074481	1.829325	2.653028	1.376706	.9061816	2	2

图 18.7　按性别和年龄进行的聚类分析结果图 6

为观测 3 类样本的特征，可以对数据进行排序操作，在主界面的“Command”文本框中输入操作命令：

```
sort _clus_2
```

并按键盘上的回车键进行确认，然后选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 18.8 所示的整理后的数据。

	V1	zv3	zv4	zv5	zv6	zv7	zv8	_clus_1	_clus_2
1	沈阳	-1.091924	-.8581841	-1.142801	-1.351635	-1.104487	-.9369602	1	1
2	郑州	-.5665381	.3700919	-.5781965	-.6271088	-1.003569	-.1088146	1	1
3	无锡	-.2199755	-.9210401	-.6844487	.3605084	-.1855144	-.4787925	1	1
4	大连	-1.277102	-.9500191	-1.267838	-1.456556	-1.221717	-1.166393	1	1
5	石家庄	-.7294003	.0526829	-.456212	-.9872448	.3010002	-.7887362	1	1
6	武汉	-.6262443	-.7007721	-.6318509	-.4212042	-.5324081	-.5667495	1	1
7	上海	-.7422578	-.8849863	-.62058	-.3641748	-.6806655	-.0099211	1	1
8	广州	-.5353548	-.9007683	-.7143872	-.4517669	-.6010191	-.542317	1	1
9	苏州	-.8294526	-.8049878	-.6040258	-1.019383	-1.432188	-.7978112	1	1
10	天津	-.6515161	-.5728834	-.6101308	-.5807921	-.9263219	-.1776909	1	1
11	杭州	-.5558974	-.6161479	-.4660741	-.7791347	.8897119	-.6563354	1	1
12	北京	-.2598782	-.1693536	-.5639904	.0425931	-.0222234	-.1658237	1	1
13	呼和浩特	.9935107	1.957137	1.327063	.6551079	-.1479303	.9811078	2	2
14	南京	1.42845	.6872288	1.369447	1.675966	1.810283	2.059628	2	2
15	长春	2.583412	1.70272	2.053453	1.502515	2.453372	2.767706	2	2
16	深圳	2.076205	-.2074481	1.829325	2.653028	1.376706	.9061816	2	2
17	长沙	.3495988	-.3147931	.1416179	.1873721	.6077508	.0843186	1	3
18	乌鲁木齐	-.0630248	2.188969	.0185768	-.1281801	-.3697567	-.211431	1	3
19	太原	.2864935	1.128853	.5870553	.0422779	.055184	-.1374355	1	3
20	哈尔滨	.3705848	.2092787	1.606724	.2519634	-.1325768	-.7033389	1	3
21	青岛	-.2836721	.630087	-.6019124	.1898928	.7728011	-.5956031	1	3
22	银川	.343983	-1.025664	.0091844	.6059555	.0935678	1.245212	1	3

图 18.8　按性别和年龄进行的聚类分析结果图 7

从图 18.8 中可以看到第 2 类所代表的人均旅游消费支出特点是无论男女老少花费支出总体上相对最高，第 1 类所代表的人均旅游消费支出特点是无论男女老少各年龄段花费支出总体上相对最低，第 3 类则表示中等水平。

图 18.9 展示的是设定聚类数为 4，然后使用“K 个平均数的聚类分析”方法进行分析的结果。在输入第 10 条 Stata 命令并且分别按键盘上的回车键进行确认后，可以看到系统产生了一个新的变量：聚类变量_clus_3（cluster name: _clus_3）。

```
. cluster kmeans zv2 zv3 zv4 zv5 zv6 zv7 zv8,k(4)
cluster name: _clus_3
```

图 18.9　按性别和年龄进行的聚类分析结果图 8

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 18.10 所示的_clus_3 数据。

	V1	zv3	zv4	zv5	zv6	zv7	zv8	_clus_1	_clus_2	_clus_3
1	沈阳	-1.091924	-.8581841	-1.142801	-1.351635	-1.104487	-.9369602	1	1	2
2	郑州	-.5665381	.3700919	-.5781965	-.6271088	-1.003569	-.1088146	1	1	1
3	无锡	-.2199755	-.9210401	-.6844487	.3605084	-.1855144	-.4787925	1	1	1
4	大连	-1.277102	-.9500191	-1.267838	-1.456556	-1.221717	-1.166393	1	1	2
5	石家庄	-.7294003	.0526829	-.456212	-.9872448	.3010002	-.7887362	1	1	1
6	武汉	-.6262443	-.7007721	-.6318509	-.4212042	-.5324081	-.5667495	1	1	1
7	上海	-.7422578	-.8849863	-.62058	-.3641748	-.6806655	-.0099211	1	1	1
8	广州	-.5353548	-.9007683	-.7143872	-.4517669	-.6010191	-.542317	1	1	1
9	苏州	-.8294526	-.8049878	-.6040258	-1.019383	-1.432188	-.7978112	1	1	2
10	天津	-.6515161	-.5728834	-.6101308	-.5807921	-.9263219	-.1776909	1	1	1
11	杭州	-.5558974	-.6161479	-.4660741	-.7791347	.8897119	-.6563354	1	1	1
12	北京	-.2598782	-.1693536	-.5639904	.0425931	-.0222234	-.1658237	1	1	1
13	呼和浩特	.9935107	1.957137	1.327063	.6551079	-.1479303	.9811078	2	2	3
14	南京	1.42845	.6872288	1.369447	1.675966	1.810283	2.059628	2	2	3
15	长春	2.583412	1.70272	2.053453	1.502515	2.453372	2.767706	2	2	3
16	深圳	2.076205	-.2074481	1.829325	2.653028	1.376706	.9061816	2	2	3
17	长沙	.3495988	-.3147931	.1416179	.1873721	.6077508	.0843186	1	3	4
18	乌鲁木齐	-.0630248	2.188969	.0185768	-.1281801	-.3697567	-.211431	1	3	4
19	太原	.2864935	1.128853	.5870553	.0422779	.055184	-.1374355	1	3	4
20	哈尔滨	.3705848	.2092787	1.606724	.2519634	-.1325768	-.7033389	1	3	4
21	青岛	-.2836721	.630087	-.6019124	.1898928	.7728011	-.5956031	1	3	4
22	银川	.343983	-1.025664	.0091844	.6059555	.0935678	1.245212	1	3	4

图 18.10　按性别和年龄进行的聚类分析结果图 9

从图 18.10 中可以看到所有的观测样本被分为 4 类，其中长春、南京、呼和浩特、深圳属于第 3 类，沈阳、大连、苏州属于第 2 类，青岛、长沙、银川、乌鲁木齐、太原、哈尔滨属于第 4 类，其他城市属于第 1 类。从图 18.9 中很难看出各个类别的特征，可以对数据进行排序操作，在主界面的“Command”文本框中输入操作命令：

```
sort _clus_3
```

并按键盘上的回车键进行确认，然后选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 18.11 所示的整理后的数据。

从图 18.11 中我们可以看出，第 3 类所代表的人均旅游消费支出特点是无论男女老少花费支出总体上相对最高，第 2 类所代表的人均旅游消费支出特点是无论男女老少各年龄段花费支出总体上相对最低，第 4 类所代表的人均旅游消费支出特点是无论男女老少各年龄段花费支出总体上相对较高，第 1 类所代表的人均旅游消费支出特点是无论男女老少各年龄段花费支出总体上相对较低。

在前面的章节中也提到过，划分聚类分析的特点是需要事先制定拟分类的数量。究竟分成多少类是合理的，这是没有定论的。用户需要根据自己的研究和需要及数据的实际特点加入自己的判断。在上面的分析中，我们尝试着把样本分别分为 2、3、4 类进行了研究，可以看出把数据分成两类是过于粗糙的，而且两个类别所包含的样本数量的差别也是比较大的，而把数据分成 3 类是比较合适的。读者可以再把数据分成 5 类、6 类或者其他数量的类别进行研究，观察分类情况，找出自己认为最优的分类。

	v1	zv3	zv4	zv5	zv6	zv7	zv8	_clus_1	_clus_2	_clus_3
1	杭州	-.5558974	-.6161479	-.4660741	-.7791347	.8897119	-.6563354	1	1	1
2	北京	-.2598782	-.1693536	-.5639904	.0425931	-.0222234	-.1658237	1	1	1
3	上海	-.7422578	-.8849863	-.62058	-.3641748	-.6806655	-.0099211	1	1	1
4	郑州	-.5665381	.3700919	-.5781965	-.6271088	-1.003569	-.1088146	1	1	1
5	武汉	-.6262443	-.7007721	-.6318509	-.4212042	-.5324081	-.5667495	1	1	1
6	石家庄	-.7294003	.0526829	-.456212	-.9872448	.3010002	-.7887362	1	1	1
7	无锡	-.2199755	-.9210401	-.6844487	.3605084	-.1855144	-.4787925	1	1	1
8	广州	-.5353548	-.9007683	-.7143872	-.4517669	-.6010191	-.542317	1	1	1
9	天津	-.6515161	-.5728834	-.6101308	-.5807921	-.9263219	-.1776909	1	1	1
10	沈阳	-1.091924	-.8581841	-1.142801	-1.351635	-1.104487	-.9369602	1	1	2
11	大连	-1.277102	-.9500191	-1.267838	-1.456556	-1.221717	-1.166393	1	1	2
12	苏州	-.8294526	-.8049878	-.6040258	-1.019383	-1.432188	-.7978112	1	1	2
13	南京	1.42845	.6872288	1.369447	1.675966	1.810283	2.059628	2	2	3
14	呼和浩特	.9935107	1.957137	1.327063	.6551079	-.1479303	.9811078	2	2	3
15	深圳	2.076205	-.2074481	1.829325	2.653028	1.376706	.9061816	2	2	3
16	长春	2.583412	1.70272	2.053453	1.502515	2.453372	2.767706	2	2	3
17	银川	.343983	-1.025664	.0091844	.6059555	.0935678	1.245212	1	3	4
18	乌鲁木齐	-.0630248	2.188969	.0185768	-.1281801	-.3697567	-.211431	1	3	4
19	哈尔滨	.3705848	.2092787	1.606724	.2519634	-.1325768	-.7033389	1	3	4
20	太原	.2864935	1.128853	.5870553	.0422779	.055184	-.1374355	1	3	4
21	长沙	.3495988	-.3147931	.1416179	.1873721	.6077508	.0843186	1	3	4
22	青岛	-.2836721	.630087	-.6019124	.1898928	.7728011	-.5956031	1	3	4

图 18.11　按性别和年龄进行的聚类分析结果图 10

通过聚类分析得到的研究结论是：按性别和年龄进行聚类分析时，青岛、长沙、银川、乌鲁木齐、太原、哈尔滨等城市的城镇居民无论男女老少，其 2007 年人均旅游消费支出都处于全国中档水平上；长春、南京、呼和浩特、深圳等城市的城镇居民无论男女老少，其 2007 年人均旅游消费支出都处于全国高档水平上；除以上城市之外的其他城市的城镇居民无论男女老少，其 2007 年人均旅游消费支出都处于全国低档水平上。

18.3.2　各城市国内旅游出游人均花费按职业进行的聚类分析

	下载资源:\video\chap18\…
	下载资源:\sample\chap18\案例18.2.dta

表 18.5 是 2007 年我国 22 个城市城镇居民国内旅游出游人均花费按职业进行分类的数据。

表 18.5 我国 2007 年城镇居民国内旅游出游人均花费情况统计（按职业分组）（单位：元/人）

城市	公务员	企事业管理人员	技术人员	商贸人员	工人
北京	1 887.9	1 270.8	1 091.9	1 289.4	733.4
天津	1 228.8	1 118.1	967.3	741.1	824.2
石家庄	1 241.6	926.6	628.4	686.3	813.2
太原	2 189.4	2 083.5	1 076.2	331.8	1 207.7
呼和浩特	3 381.6	2 729.6	1 945.8	2 553.1	3 077.8
沈阳	632.5	530.8	374.2	299.2	281.8
大连	1 136.9	478.1	363.0	342.8	277.5
长春	2 547.0	3 400.7	1 815.8	1 492.2	986.7
哈尔滨	2 559.3	3 403.9	1 997.3	1 484.4	845.0
上海	1 482.6	2 126.8	1 186.0	819.6	759.4
南京	3 934.3	2 259.1	2 987.7	1 985.9	1 641.2
无锡	0	1 552.2	2 398.8	1 425.6	706.2
苏州	233.2	1 114.0	218.9	518.3	401.9
杭州	2 007.1	1 378.0	987.9	728.8	673.0
青岛	1 825.2	1 155.4	1 566.6	1 407.5	1 047.7
郑州	776.9	1 551.0	1 832.5	643.3	691.3
武汉	1 113.7	996.4	1 500.3	704.7	803.5
长沙	939.2	1 877.8	1 926.2	1 022.3	995.8
广州	1 182.8	940.1	970.9	726.2	829.7
深圳	4 412.8	3 455.2	1 871.4	2 247.8	3 934.3
银川	1 448.4	2 487.0	2 133.4	1 152.6	1 465.8
乌鲁木齐	1 854.8	461.3	1 959.7	890.0	930.8

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有 6 个变量，分别为“城市”、“公务员”“企事业管理人员”“技术人员”“商贸人员”“工人”。我们将这 6 个变量分别定义为 V1~V6，然后录入相关数据。录入完成后数据如图 18.12 所示。

	V1	V2	V3	V4	V5	V6
1	北京	1787.9	1270.8	1091.9	1289.4	733.4
2	天津	1228.8	1118.1	967.3	741.1	824.2
3	石家庄	1241.6	926.6	628.4	686.3	813.2
4	太原	2189.4	2083.5	1076.2	331.8	1207.7
5	呼和浩特	3381.6	2729.6	1945.8	2553.1	3077.8
6	沈阳	632.5	530.8	374.2	299.2	281.8
7	大连	1136.9	478.1	363	342.8	277.5
8	长春	2547	3400.7	1815.8	1492.2	986.7
9	哈尔滨	2559.3	3403.9	1997.3	1484.4	845
10	上海	1482.6	2126.8	1176	819.6	759.4
11	南京	3934.3	2259.1	2987.7	1985.9	1641.2
12	无锡	0	1552.2	2398.8	1425.6	706.2
13	苏州	233.2	1114	218.9	517.3	401.9
14	杭州	2007.1	1378	987.9	728.8	673
15	青岛	1725.2	1155.4	1566.6	1407.5	1047.7
16	郑州	776.9	1551	1732.5	643.3	691.3
17	武汉	1113.7	996.4	1500.3	704.7	803.5
18	长沙	939.2	1877.8	1926.2	1022.3	995.8
19	广州	1172.8	940.1	970.9	726.2	829.7
20	深圳	4412.8	3455.2	1871.4	2247.8	3934.3
21	银川	1448.4	2487	2133.4	1152.6	1465.8
22	乌鲁木齐	1754.8	461.3	1959.7	890	930.8

图 18.12 案例 18.2 数据

聚类分析的分析步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中分别输入如下命令并按键盘上的回车键进行确认。

- egen zv2=std(V2)：本命令旨在对 V2 变量数据进行标准化处理，标准化处理方式是使变量的平均数为 0 而且标准差为 1。
- egen zv3=std(V3)：本命令旨在对 V3 变量数据进行标准化处理，标准化处理方式是使变量的平均数为 0 而且标准差为 1。
- egen zv4=std(V4)：本命令旨在对 V4 变量数据进行标准化处理，标准化处理方式是使变量的平均数为 0 而且标准差为 1。
- egen zv5=std(V5)：本命令旨在对 V5 变量数据进行标准化处理，标准化处理方式是使变量的平均数为 0 而且标准差为 1。
- egen zv6=std(V6)：本命令旨在对 V6 变量数据进行标准化处理，标准化处理方式是使变量的平均数为 0 而且标准差为 1。
- cluster kmeans zv2 zv3 zv4 zv5 zv6,k(2)：本命令旨在对 V2~V6 的标准化变量进行“K 个平均数的聚类分析”，设定的聚类数是 2。
- cluster kmeans zv2 zv3 zv4 zv5 zv6,k(3)：本命令旨在对 V2~V6 的标准化变量进行“K 个平均数的聚类分析”，设定的聚类数是 3。
- cluster kmeans zv2 zv3 zv4 zv5 zv6,k(4)：本命令旨在对 V2~V6 的标准化变量进行“K 个平均数的聚类分析”，设定的聚类数是 4。

03 设置完毕后，按键盘上的回车键，等待输出结果。

在 Stata 14.0 主界面的结果窗口可以看到如图 18.13~图 18.22 所示的分析结果。

1．数据标准化处理

在分析过程中前 5 条 Stata 命令旨在对数据进行标准化处理，选择的标准化处理方式是使变量的平均数为 0 而且标准差为 1。之所以这样做是因为我们进行聚类分析的变量都是以不可比的单位进行的测度，它们具有极为不同的方差，对数据进行标准化处理可以避免使结果受到具有最大方差变量的影响。在输入前 5 条 Stata 命令并且分别按键盘上的回车键进行确认后，选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 18.13 所示的变换后的数据。

	v5	v6	zv2	zv3	zv4	zv5	zv6
1	1289.4	733.4	.0664084	-.4509933	-.487912	.3602029	-.413554
2	741.1	824.2	-.4353952	-.613227	-.6623245	-.5310941	-.307548
3	686.3	813.2	-.423907	-.816683	-1.136709	-.620175	-.3203902
4	331.8	1207.7	.4267627	.4124463	-.5098886	-1.196438	.1401756
5	2553.1	3077.8	1.496787	1.098885	.7073593	2.414429	2.323456
6	299.2	281.8	-.9705867	-1.237194	-1.492533	-1.249431	-.9407823
7	342.8	277.5	-.5178774	-1.293184	-1.508211	-1.178556	-.9458025
8	1492.2	986.7	.7477161	1.811884	.525388	.6898673	-.1178346
9	1484.4	845	.7587556	1.815283	.7794479	.677188	-.2832647
10	819.6	759.4	-.2076046	.4584498	-.3701906	-.4034873	-.3831999
11	1985.9	1641.2	1.992846	.5990098	2.165789	1.492409	.6462727
12	1425.6	706.2	-1.538268	-.1520246	1.341459	.5816047	-.4453092
13	517.3	401.9	-1.328967	-.6175829	-1.709919	-.8948954	-.8005695
14	728.8	673	.2631448	-.3371005	-.633489	-.5510885	-.4840691
15	1407.5	1047.7	.0101338	-.5735981	.176563	.552182	-.0466192
16	643.3	691.3	-.8409847	-.1532995	.4087863	-.6900743	-.4627045
17	704.7	803.5	-.5386999	-.7425251	.0837577	-.5902646	-.3317146
18	1022.3	995.8	-.6953171	.1939038	.6799234	-.0739854	-.1072107
19	726.2	829.7	-.4856564	-.8023401	-.6572852	-.555315	-.301127
20	2247.8	3934.3	2.422309	1.869786	.6032156	1.918144	3.323392
21	1152.6	1465.8	-.2382998	.8411382	.9699575	.1378257	.4414991
22	890	930.8	.0367005	-1.311033	.726816	-.2890475	-.183096

图 18.13　按职业进行的聚类分析结果图 1

2．K 个平均数的聚类分析

图 18.14 展示的是设定聚类数为 2，然后使用“K 个平均数的聚类分析”方法进行分析的结果。在输入第 6 条 Stata 命令并且分别按键盘上的回车键进行确认后，我们可以看到系统产生了一个新的聚类变量。

```
. cluster kmeans zv2 zv3 zv4 zv5 zv6,k(2)
cluster name: _clus_1
```

图 18.14　按职业进行的聚类分析结果图 2

选择“Data” | “Data Editor” | “Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 18.15 所示的聚类数据。

	v1	zv2	zv3	zv4	zv5	zv6	_clus_1
1	北京	.0664084	-.4509933	-.487912	.3602029	-.413554	1
2	天津	-.4353952	-.613227	-.6623245	-.5310941	-.307548	1
3	石家庄	-.423907	-.816683	-1.136709	-.620175	-.3203902	1
4	太原	.4267627	.4124463	-.5098886	-1.196438	.1401756	1
5	呼和浩特	1.496787	1.098885	.7073593	2.414429	2.323456	2
6	沈阳	-.9705867	-1.237194	-1.492533	-1.249431	-.9407823	1
7	大连	-.5178774	-1.293184	-1.508211	-1.178556	-.9458025	1
8	长春	.7477161	1.811884	.525388	.6898673	-.1178346	2
9	哈尔滨	.7587556	1.815283	.7794479	.677188	-.2832647	2
10	上海	-.2076046	.4584498	-.3701906	-.4034873	-.3831999	1
11	南京	1.992846	.5990098	2.165789	1.492409	.6462727	2
12	无锡	-1.538268	-.1520246	1.341459	.5816047	-.4453092	1
13	苏州	-1.328967	-.6175829	-1.709919	-.8948954	-.8005695	1
14	杭州	.2631448	-.3371005	-.633489	-.5510885	-.4840691	1
15	青岛	.0101338	-.5735981	.176563	.552182	-.0466192	1
16	郑州	-.8409847	-.1532995	.4087863	-.6900743	-.4627045	1
17	武汉	-.5386999	-.7425251	.0837577	-.5902646	-.3317146	1
18	长沙	-.6953171	.1939038	.6799234	-.0739854	-.1072107	1
19	广州	-.4856564	-.8023401	-.6572852	-.555315	-.301127	1
20	深圳	2.422309	1.869786	.6032156	1.918144	3.323392	2
21	银川	-.2382998	.8411382	.9699575	.1378257	.4414991	2
22	乌鲁木齐	.0367005	-1.311033	.726816	-.2890475	-.183096	1

图 18.15　按职业进行的聚类分析结果图 3

从图 18.15 中可以看到所有的观测样本被分为两类，其中呼和浩特、哈尔滨、长春、南京、深圳、银川被分到第 2 类，其他的省市被分到第 1 类。

为观测两类样本的特征，可以对数据进行排序操作，在主界面的“Command”文本框中输入操作命令：

```
sort _clus_1
```

并按键盘上的回车键进行确认，然后选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 18.16 所示的整理后的数据。

可以看到第 2 类所代表的人均旅游消费支出的特点是无论职业类型如何花费支出总体上相对较高，第 1 类所代表的人均旅游消费支出特点是无论职业类型如何花费支出总体上相对较低。

图 18.17 展示的是设定聚类数为 3，然后使用“K 个平均数的聚类分析”方法进行分析的结果。在输入第 7 条 Stata 命令并且分别按键盘上的回车键进行确认后，可以看到系统产生了一个新的变量：聚类变量_clus_2（cluster name: _clus_2）。

	v1	zv2	zv3	zv4	zv5	zv6	_clus_1
1	上海	-.2076046	.4584498	-.3701906	-.4034873	-.3831999	1
2	长沙	-.6953171	.1939038	.6799234	-.0739854	-.1072107	1
3	郑州	-.8409847	-.1532995	.4087863	-.6900743	-.4627045	1
4	北京	.0664084	-.4509933	-.487912	.3602029	-.413554	1
5	沈阳	-.9705867	-1.237194	-1.492533	-1.249431	-.9407823	1
6	杭州	.2631448	-.3371005	-.633489	-.5510885	-.4840691	1
7	青岛	.0101338	-.5735981	.176563	.552182	-.0466192	1
8	武汉	-.5386999	-.7425251	.0837577	-.5902646	-.3317146	1
9	天津	-.4353952	-.613227	-.6623245	-.5310941	-.307548	1
10	乌鲁木齐	.0367005	-1.311033	.726816	-.2890475	-.183096	1
11	石家庄	-.423907	-.816683	-1.136709	-.620175	-.3203902	1
12	广州	-.4856564	-.8023401	-.6572852	-.555315	-.301127	1
13	太原	.4267627	.4124463	-.5098886	-1.196438	.1401756	1
14	大连	-.5178774	-1.293184	-1.508211	-1.178556	-.9458025	1
15	苏州	-1.328967	-.6175829	-1.709919	-.8948954	-.8005695	1
16	无锡	-1.538268	-.1520246	1.341459	.5816047	-.4453092	1
17	银川	-.2382998	.8411382	.9699575	.1378257	.4414991	2
18	长春	.7477161	1.811884	.525388	.6898673	-.1178346	2
19	哈尔滨	.7587556	1.815283	.7794479	.677188	-.2832647	2
20	南京	1.992846	.5990098	2.165789	1.492409	.6462727	2
21	呼和浩特	1.496787	1.098885	.7073593	2.414429	2.323456	2
22	深圳	2.422309	1.869786	.6032156	1.918144	3.323392	2

图 18.16　按职业进行的聚类分析结果图 4

```
. cluster kmeans zv2 zv3 zv4 zv5 zv6,k(3)
cluster name:  _clus_2
```

图 18.17　按职业进行的聚类分析结果图 5

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 18.18 所示的_clus_2 数据。

	v1	zv3	zv4	zv5	zv6	zv7	zv8	_clus_1	_clus_2
1	青岛	-.2836721	.630087	-.6019124	.1898928	.7728011	-.5956031	1	3
2	杭州	-.5558974	-.6161479	-.4660741	-.7791347	.8897119	-.6563354	1	1
3	沈阳	-1.091924	-.8581841	-1.142801	-1.351635	-1.104487	-.9369602	1	1
4	北京	-.2598782	-.1693536	-.5639904	.0425931	-.0222234	-.1658237	1	1
5	郑州	-.5665381	.3700919	-.5781965	-.6271088	-1.003569	-.1088146	1	1
6	长沙	.3495988	-.3147931	.1416179	.1873721	.6077508	.0843186	1	3
7	上海	-.7422578	-.8849863	-.62058	-.3641748	-.6806655	-.0099211	1	1
8	银川	.343983	-1.025664	.0091844	.6059555	.0935678	1.245212	1	3
9	武汉	-.6262443	-.7007721	-.6318509	-.4212042	-.5324081	-.5667495	1	1
10	乌鲁木齐	-.0630248	2.188969	.0185768	-.1281801	-.3697567	-.211431	1	3
11	苏州	-.8294526	-.8049878	-.6040258	-1.019383	-1.432188	-.7978112	1	1
12	广州	-.5353548	-.9007683	-.7143872	-.4517669	-.6010191	-.542317	1	1
13	石家庄	-.7294003	.0526829	-.456212	-.9872448	.3010002	-.7887362	1	1
14	大连	-1.277102	-.9500191	-1.267838	-1.456556	-1.221717	-1.166393	1	1
15	无锡	-.2199755	-.9210401	-.6844487	.3605084	-.1855144	-.4787925	1	1
16	太原	.2864935	1.128853	.5870553	.0422779	.055184	-.1374355	1	3
17	天津	-.6515161	-.5728834	-.6101308	-.5807921	-.9263219	-.1776909	1	1
18	哈尔滨	.3705848	.2092787	1.606724	.2519634	-.1325768	-.7033389	1	3
19	长春	2.583412	1.70272	2.053453	1.502515	2.453372	2.767706	2	2
20	南京	1.42845	.6872288	1.369447	1.675966	1.810283	2.059628	2	2
21	呼和浩特	.9935107	1.957137	1.327063	.6551079	-.1479303	.9811078	2	2
22	深圳	2.076205	-.2074481	1.829325	2.653028	1.376706	.9061816	2	2

图 18.18　按职业进行的聚类分析结果图 6

从图 18.18 中可以看到所有的观测样本被分为 3 类，其中长春、南京、呼和浩特、深圳属于第 2 类，上海、郑州、北京、杭州、武汉、青岛、长沙、银川、乌鲁木齐、太原、哈尔滨、无锡属于第 3 类，其他城市属于第 1 类。

为观测 3 类样本的特征，我们可以对数据进行排序操作，在主界面的“Command”文本框中输入操作命令：

```
sort _clus_2
```

并按键盘上的回车键进行确认，然后选择“Data” | “Data Editor” | “Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 18.19 所示的整理后的数据。

	v1	zv2	zv3	zv4	zv5	zv6	_clus_1	_clus_2
1	天津	-.4353952	-.613227	-.6623245	-.5310941	-.307548	1	1
2	大连	-.5178774	-1.293184	-1.508211	-1.178556	-.9458025	1	1
3	苏州	-1.328967	-.6175829	-1.709919	-.8948954	-.8005695	1	1
4	沈阳	-.9705867	-1.237194	-1.492533	-1.249431	-.9407823	1	1
5	广州	-.4856564	-.8023401	-.6572852	-.555315	-.301127	1	1
6	石家庄	-.423907	-.816683	-1.136709	-.620175	-.3203902	1	1
7	呼和浩特	1.496787	1.098885	.7073593	2.414429	2.323456	2	2
8	深圳	2.422309	1.869786	.6032156	1.918144	3.323392	2	2
9	南京	1.992846	.5990098	2.165789	1.492409	.6462727	2	2
10	长春	.7477161	1.811884	.525388	.6898673	-.1178346	2	2
11	杭州	.2631448	-.3371005	-.633489	-.5510885	-.4840691	1	3
12	郑州	-.8409847	-.1532995	.4087863	-.6900743	-.4627045	1	3
13	哈尔滨	.7587556	1.815283	.7794479	.677188	-.2832647	2	3
14	北京	.0664084	-.4509933	-.487912	.3602029	-.413554	1	3
15	长沙	-.6953171	.1939038	.6799234	-.0739854	-.1072107	1	3
16	银川	-.2382998	.8411382	.9699575	.1378257	.4414991	2	3
17	太原	.4267627	.4124463	-.5098886	-1.196438	.1401756	1	3
18	上海	-.2076046	.4584498	-.3701906	-.4034873	-.3831999	1	3
19	武汉	-.5386999	-.7425251	.0837577	-.5902646	-.3317146	1	3
20	乌鲁木齐	.0367005	-1.311033	.726816	-.2890475	-.183096	1	3
21	青岛	.0101338	-.5735981	.176563	.552182	-.0466192	1	3
22	无锡	-1.538268	-.1520246	1.341459	.5816047	-.4453092	1	3

图 18.19 按职业进行的聚类分析结果图 7

从图 18.19 中可以看到第 2 类所代表的人均旅游消费支出特点是无论职业类型如何花费支出总体上相对最高，第 1 类所代表的人均旅游消费支出特点是无论职业类型如何花费支出总体上相对最低，第 3 类则表示中等水平。

图 18.20 展示的是设定聚类数为 4，然后使用“K 个平均数的聚类分析”方法进行分析的结果。在输入第 8 条 Stata 命令并且分别按键盘上的回车键进行确认后，可以看到系统产生了一个新的变量：聚类变量_clus_3（cluster name: _clus_3）。

```
. cluster kmeans zv2 zv3 zv4 zv5 zv6,k(4)
cluster name: _clus_3
```

图 18.20 按职业进行的聚类分析结果图 8

选择“Data” | “Data Editor” | “Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 18.21 所示的_clus_3 数据。

	V1	zv2	zv3	zv4	zv5	zv6	_clus_1	_clus_2	_clus_3
1	天津	-.4353952	-.613227	-.6623245	-.5310941	-.307548	1	1	4
2	大连	-.5178774	-1.293184	-1.508211	-1.178556	-.9458025	1	1	2
3	苏州	-1.328967	-.6175829	-1.709919	-.8948954	-.8005695	1	1	1
4	沈阳	-.9705867	-1.237194	-1.492533	-1.249431	-.9407823	1	1	2
5	广州	-.4856564	-.8023401	-.6572852	-.555315	-.301127	1	1	4
6	石家庄	-.423907	-.816683	-1.136709	-.620175	-.3203902	1	1	2
7	呼和浩特	1.496787	1.098885	.7073593	2.414429	2.323456	2	2	3
8	深圳	2.422309	1.869786	.6032156	1.918144	3.323392	2	2	3
9	南京	1.992846	.5990098	2.165789	1.492409	.6462727	2	2	3
10	长春	.7477161	1.811884	.525388	.6898673	-.1178346	2	2	3
11	杭州	.2631448	-.3371005	-.633489	-.5510885	-.4840691	1	3	4
12	郑州	-.8409847	-.1532995	.4087863	-.6900743	-.4627045	1	3	4
13	哈尔滨	.7587556	1.815283	.7794479	.677188	-.2832647	2	3	3
14	北京	.0664084	-.4509933	-.487912	.3602029	-.413554	1	3	4
15	长沙	-.6953171	.1939038	.6799234	-.0739854	-.1072107	1	3	4
16	银川	-.2382998	.8411382	.9699575	.1378257	.4414991	2	3	4
17	太原	.4267627	.4124463	-.5098886	-1.196438	.1401756	1	3	4
18	上海	-.2076046	.4584498	-.3701906	-.4034873	-.3831999	1	3	4
19	武汉	-.5386999	-.7425251	.0837577	-.5902646	-.3317146	1	3	4
20	乌鲁木齐	.0367005	-1.311033	.726816	-.2890475	-.183096	1	3	4
21	青岛	.0101338	-.5735981	.176563	.552182	-.0466192	1	3	4
22	无锡	-1.538268	-.1520246	1.341459	.5816047	-.4453092	1	3	4

图 18.21　按职业进行的聚类分析结果图 9

从图 18.21 中可以看到所有的观测样本被分为 4 类，其中大连、沈阳、石家庄属于第 2 类，长春、南京、呼和浩特、深圳、哈尔滨属于第 3 类，苏州属于第 1 类，其他城市属于第 4 类。从图 18.20 中很难看出各个类别的特征，可以对数据进行排序操作，在主界面的“Command”文本框中输入操作命令：

```
sort _clus_3
```

并按键盘上的回车键进行确认，然后选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 18.22 所示的整理后的数据。

	V1	zv2	zv3	zv4	zv5	zv6	_clus_1	_clus_2	_clus_3
1	苏州	-1.328967	-.6175829	-1.709919	-.8948954	-.8005695	1	1	1
2	石家庄	-.423907	-.816683	-1.136709	-.620175	-.3203902	1	1	2
3	大连	-.5178774	-1.293184	-1.508211	-1.178556	-.9458025	1	1	2
4	沈阳	-.9705867	-1.237194	-1.492533	-1.249431	-.9407823	1	1	2
5	呼和浩特	1.496787	1.098885	.7073593	2.414429	2.323456	2	2	3
6	哈尔滨	.7587556	1.815283	.7794479	.677188	-.2832647	2	3	3
7	长春	.7477161	1.811884	.525388	.6898673	-.1178346	2	2	3
8	深圳	2.422309	1.869786	.6032156	1.918144	3.323392	2	2	3
9	南京	1.992846	.5990098	2.165789	1.492409	.6462727	2	2	3
10	上海	-.2076046	.4584498	-.3701906	-.4034873	-.3831999	1	3	4
11	广州	-.4856564	-.8023401	-.6572852	-.555315	-.301127	1	1	4
12	北京	.0664084	-.4509933	-.487912	.3602029	-.413554	1	3	4
13	无锡	-1.538268	-.1520246	1.341459	.5816047	-.4453092	1	3	4
14	郑州	-.8409847	-.1532995	.4087863	-.6900743	-.4627045	1	3	4
15	天津	-.4353952	-.613227	-.6623245	-.5310941	-.307548	1	1	4
16	杭州	.2631448	-.3371005	-.633489	-.5510885	-.4840691	1	3	4
17	长沙	-.6953171	.1939038	.6799234	-.0739854	-.1072107	1	3	4
18	乌鲁木齐	.0367005	-1.311033	.726816	-.2890475	-.183096	1	3	4
19	银川	-.2382998	.8411382	.9699575	.1378257	.4414991	2	3	4
20	武汉	-.5386999	-.7425251	.0837577	-.5902646	-.3317146	1	3	4
21	太原	.4267627	.4124463	-.5098886	-1.196438	.1401756	1	3	4
22	青岛	.0101338	-.5735981	.176563	.552182	-.0466192	1	3	4

图 18.22　按职业进行的聚类分析结果图 10

从图 18.22 中可以看到第 3 类所代表的人均旅游消费支出特点是无论职业类型如何花费支出总体上相对最高，第 2 类所代表的人均旅游消费支出特点是无论职业类型如何花费支出总体上相对较低，第 4 类所代表的人均旅游消费支出特点是无论职业类型如何花费支出总体上相对较高，第 1 类所代表的人均旅游消费支出特点是无论职业类型如何花费支出总体上相对最低。

在前面的章节中也提到过，划分聚类分析的特点是需要事先制定拟分类的数量。究竟分成多少类是合理的，这是没有定论的。用户需要根据自己的研究和需要及数据的实际特点加入

自己的判断。在上面的分析中，我们尝试着把样本分别分为 2、3、4 类进行了研究，可以看出把数据分成两类是过于粗糙的，而且两个类别所包含的样本数量的差别也是比较大的，而把数据分成 3 类是比较合适的。读者可以再把数据分成 5 类、6 类或者其他数量的类别进行研究，观察分类情况，找出自己认为的最优分类。

通过聚类分析得到的研究结论是：按职业进行聚类分析时，上海、郑州、北京、杭州、武汉、青岛、长沙、银川、乌鲁木齐、太原、哈尔滨、无锡等城市的城镇居民无论职业类型如何，其 2007 年人均旅游消费支出都处于全国中档水平上；长春、南京、呼和浩特、深圳等城市的城镇居民无论职业类型如何，其 2007 年人均旅游消费支出都处于全国高档水平上；除以上城市之外的其他城市的城镇居民无论职业类型如何，其 2007 年人均旅游消费支出都处于全国低档水平上。

18.3.3　各城市国内旅游出游人均花费按文化水平进行的聚类分析

	下载资源:\video\chap18\…
	下载资源:\sample\chap18\案例18.3.dta

表 18.6 是 2007 年我国 22 个城市城镇居民国内旅游出游人均花费按文化水平进行分类的数据。

表 18.6　我国 2007 年城镇居民国内旅游出游人均花费情况统计（按文化水平分组）（单位：元/人）

城市	大专及以上	中专及高中	初中	小学	小学以下
北京	1 322.4	868.9	757.8	585.8	355.1
天津	891.9	826.1	829.3	509.0	433.0
石家庄	978.7	855.2	501.8	580.6	486.6
太原	1 634.2	1 866.9	979.6	1 180.4	275.0
呼和浩特	2 378.6	1 926.3	1 600.3	248.4	1 686.7
沈阳	501.3	367.5	319.3	394.8	391.8
大连	433.3	321.2	213.0	188.9	261.2
长春	2 909.3	1 385.6	1 200.8	1 864.5	0
哈尔滨	2 561.6	1 857.5	950.8	295.8	470.8
上海	1 082.8	1 098.3	425.6	567.2	699.3
南京	2 647.9	1 986.2	1 933.0	2 244.0	1 211.4
无锡	1 519.3	1 030.2	410.4	385.0	562.8
苏州	898.7	501.0	694.7	520.7	420.6
杭州	1 224.7	771.4	866.8	557.6	418.3
青岛	1 352.2	1 140.8	327.4	831.9	1 432.8
郑州	933.0	882.2	880.0	559.2	316.8
武汉	1 139.8	683.2	818.0	704.2	421.0
长沙	1 569.2	1 319.6	667.4	1 187.2	87.3
广州	1 066.2	746.7	787.5	500.2	394.8
深圳	3 256.3	2 464.7	1 868.2	1 474.8	1 321.3
银川	1 890.7	1 403.4	895.5	1 670.4	189.3
乌鲁木齐	1 808.3	776.0	489.7	580.5	540.9

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有 6 个变量，分别为“城市”“大专及以上”“中专及高中”“初中”“小学”“小学以下”。我们将这 6 个变量分别定义为 V1~V6，然后录入相关数据。录入完成后数据如图 18.23 所示。

	V1	V2	V3	V4	V5	V6
1	北京	1322.4	868.9	757.8	585.8	355.1
2	天津	891.9	826.1	829.3	509	433
3	石家庄	978.7	855.2	501.8	580.6	486.6
4	太原	1634.2	1866.9	979.6	1170.4	275
5	呼和浩特	2378.6	1926.3	1600.3	248.4	1686.7
6	沈阳	501.3	367.5	319.3	394.8	391.8
7	大连	433.3	321.2	213	178.9	261.2
8	长春	2909.3	1385.6	1200.8	1864.5	0
9	哈尔滨	2561.6	1857.5	950.8	295.8	470.8
10	上海	1082.8	1098.3	425.6	567.2	699.3
11	南京	2647.9	1986.2	1933	2244	1211.4
12	无锡	1519.3	1030.2	410.4	385	562.8
13	苏州	898.7	501	694.7	520.7	420.6
14	杭州	1224.7	771.4	866.8	557.6	418.3
15	青岛	1352.2	1140.8	327.4	831.9	1432.8
16	郑州	933	882.2	880	559.2	316.8
17	武汉	1139.8	683.2	817	704.2	421
18	长沙	1569.2	1319.6	667.4	1177.2	87.3
19	广州	1066.2	746.7	787.5	500.2	394.8
20	深圳	3256.3	2464.7	1868.2	1474.8	1321.3
21	银川	1890.7	1403.4	895.5	1670.4	179.3
22	乌鲁木齐	1708.3	776	489.7	580.5	540.9

图 18.23　案例 18.3 数据

聚类分析的分析步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中分别输入如下命令并按键盘上的回车键进行确认。

- egen zv2=std(V2)：本命令旨在对 V2 变量数据进行标准化处理，标准化处理方式是使变量的平均数为 0 而且标准差为 1。
- egen zv3=std(V3)：本命令旨在对 V3 变量数据进行标准化处理，标准化处理方式是使变量的平均数为 0 而且标准差为 1。
- egen zv4=std(V4)：本命令旨在对 V4 变量数据进行标准化处理，标准化处理方式是使变量的平均数为 0 而且标准差为 1。
- egen zv5=std(V5)：本命令旨在对 V5 变量数据进行标准化处理，标准化处理方式是使变量的平均数为 0 而且标准差为 1。
- egen zv6=std(V6)：本命令旨在对 V6 变量数据进行标准化处理，标准化处理方式是使变量的平均数为 0 而且标准差为 1。
- cluster kmeans zv2 zv3 zv4 zv5 zv6,k(2)：本命令旨在对 V2~V6 的标准化变量进行“K 个平均数的聚类分析”，设定的聚类数是 2。
- cluster kmeans zv2 zv3 zv4 zv5 zv6,k(3)：本命令旨在对 V2~V6 的标准化变量进行“K 个平均数的聚类分析”，设定的聚类数是 3。
- cluster kmeans zv2 zv3 zv4 zv5 zv6,k(4)：本命令旨在对 V2~V6 的标准化变量进行“K 个平均数的聚类分析”，设定的聚类数是 4。

03 设置完毕后，按键盘上的回车键，等待输出结果。

在 Stata 14.0 主界面的结果窗口可以看到如图 18.24~图 18.33 所示的分析结果。

1．数据标准化处理

在分析过程中前 5 条 Stata 命令旨在对数据进行标准化处理，选择的标准化处理方式是使变量的平均数为 0 而且标准差为 1，之所以这样做是因为我们进行聚类分析的变量都是以不可比的单位进行的测度，它们具有极为不同的方差，对数据进行标准化处理可以避免使结果受到具有最大方差变量的影响。在输入前 5 条 Stata 命令并且分别按键盘上的回车键进行确认后，我们选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 18.24 所示的变换后的数据。

	v5	v6	zv2	zv3	zv4	zv5	zv6
1	585.8	355.1	-.2826713	-.4716702	-.1698035	-.3831225	-.4655777
2	509	433	-.8395354	-.5461491	-.0166757	-.5204565	-.2903906
3	580.6	486.6	-.7272571	-.4955104	-.7180651	-.3924212	-.1698511
4	1170.4	275	.1206509	1.265009	.3052139	.662261	-.6457124
5	248.4	1686.7	1.083554	1.368374	1.634534	-.9864622	2.529019
6	394.8	391.8	-1.344788	-1.344186	-1.108916	-.7246693	-.3830442
7	178.9	261.2	-1.432748	-1.424755	-1.336573	-1.110742	-.6767467
8	1864.5	0	1.770029	.4274701	.7789464	1.903453	-1.264152
9	295.8	470.8	1.320269	1.248651	.2435345	-.9017014	-.2053833
10	567.2	699.3	-.5926007	-.0724776	-.8812586	-.416383	.3084838
11	2244	1211.4	1.431901	1.47261	2.34706	2.582076	1.460131
12	385	562.8	-.0279754	-.1909826	-.9138117	-.7421936	.0015128
13	520.7	420.6	-.8307394	-1.111875	-.3049413	-.4995345	-.3182766
14	557.6	418.3	-.4090491	-.6413357	.0636361	-.4335499	-.3234491
15	831.9	1432.8	-.2441242	.0014792	-1.091568	.0569543	1.958031
16	559.2	316.8	-.7863714	-.4485261	.0919059	-.4306886	-.5517095
17	704.2	421	-.5188696	-.7948178	-.0430179	-.1713992	-.3173771
18	1177.2	87.3	.0365715	.3126196	-.3634083	.6744207	-1.067826
19	500.2	394.8	-.6140734	-.6843176	-.1061965	-.5361927	-.3762976
20	1474.8	1321.3	2.218884	2.305276	2.208282	1.20659	1.707282
21	1670.4	179.3	.452441	.458445	.1251014	1.556363	-.8609294
22	580.5	540.9	.2165015	-.633331	-.743979	-.3926	-.0477374

图 18.24　按文化水平进行的聚类分析结果图 1

2．K 个平均数的聚类分析

图 18.25 展示的是设定聚类数为 2，然后使用“K 个平均数的聚类分析”方法进行分析的结果。在输入第 6 条 Stata 命令并且分别按键盘上的回车键进行确认后，可以看到系统产生了一个新的聚类变量。

```
. cluster kmeans zv2 zv3 zv4 zv5 zv6,k(2)
cluster name: _clus_1
```

图 18.25　按文化水平进行的聚类分析结果图 2

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 18.26 所示的聚类数据。

	v1	zv2	zv3	zv4	zv5	zv6	_clus_1
1	北京	-.2826713	-.4716702	-.1698035	-.3831225	-.4655777	1
2	天津	-.8395354	-.5461491	-.0166757	-.5204565	-.2903906	1
3	石家庄	-.7272571	-.4955104	-.7180651	-.3924212	-.1698511	1
4	太原	.1206509	1.265009	.3052139	.662261	-.6457124	2
5	呼和浩特	1.083554	1.368374	1.634534	-.9864622	2.529019	2
6	沈阳	-1.344788	-1.344186	-1.108916	-.7246693	-.3830442	1
7	大连	-1.432748	-1.424755	-1.336573	-1.110742	-.6767467	1
8	长春	1.770029	.4274701	.7789464	1.903453	-1.264152	2
9	哈尔滨	1.320269	1.248651	.2435345	-.9017014	-.2053833	2
10	上海	-.5926007	-.0724776	-.8812586	-.416383	.3084838	1
11	南京	1.431901	1.47261	2.34706	2.582076	1.460131	2
12	无锡	-.0279754	-.1909826	-.9138117	-.7421936	.0015128	1
13	苏州	-.8307394	-1.111875	-.3049413	-.4995345	-.3182766	1
14	杭州	-.4090491	-.6413357	.0636361	-.4335499	-.3234491	1
15	青岛	-.2441242	.0014792	-1.091568	.0569543	1.958031	1
16	郑州	-.7863714	-.4485261	.0919059	-.4306886	-.5517095	1
17	武汉	-.5188696	-.7948178	-.0430179	-.1713992	-.3173771	1
18	长沙	.0365715	.3126196	-.3634083	.6744207	-1.067826	1
19	广州	-.6140734	-.6843176	-.1061965	-.5361927	-.3762976	1
20	深圳	2.218884	2.305276	2.208282	1.20659	1.707282	2
21	银川	.452441	.458445	.1251014	1.556363	-.8609294	2
22	乌鲁木齐	.2165015	-.633331	-.743979	-.3926	-.0477374	1

图 18.26　按文化水平进行的聚类分析结果图 3

从图 18.26 中可以看到所有的观测样本被分为两类，其中太原、呼和浩特、哈尔滨、长春、南京、深圳、银川被分到第 2 类，其他的省市被分到第 1 类。

为观测两类样本的特征，可以对数据进行排序操作，在主界面的“Command”文本框中输入操作命令：

```
sort _clus_1
```

并按键盘上的回车键进行确认，然后选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 18.27 所示的整理后的数据。

	V1	zv2	zv3	zv4	zv5	zv6	_clus_1
1	上海	-.5926007	-.0724776	-.8812586	-.416383	.3084838	1
2	长沙	.0365715	.3126196	-.3634083	.6744207	-1.067826	1
3	郑州	-.7863714	-.4485261	.0919059	-.4306886	-.5517095	1
4	北京	-.2826713	-.4716702	-.1698035	-.3831225	-.4655777	1
5	沈阳	-1.344788	-1.344186	-1.108916	-.7246693	-.3830442	1
6	杭州	-.4090491	-.6413357	.0636361	-.4335499	-.3234491	1
7	青岛	-.2441242	.0014792	-1.091568	.0569543	1.958031	1
8	武汉	-.5188696	-.7948178	-.0430179	-.1713992	-.3173771	1
9	天津	-.8395354	-.5461491	-.0166757	-.5204565	-.2903906	1
10	乌鲁木齐	.2165015	-.633331	-.743979	-.3926	-.0477374	1
11	石家庄	-.7272571	-.4955104	-.7180651	-.3924212	-.1698511	1
12	广州	-.6140734	-.6843176	-.1061965	-.5361927	-.3762976	1
13	大连	-1.432748	-1.424755	-1.336573	-1.110742	-.6767467	1
14	苏州	-.0307394	-1.111075	-.3049413	-.4995345	-.3102766	1
15	无锡	-.0279754	-.1909826	-.9138117	-.7421936	.0015128	1
16	银川	.452441	.458445	.1251014	1.556363	-.8609294	2
17	太原	.1206509	1.265009	.3052139	.662261	-.6457124	2
18	长春	1.770029	.4274701	.7789464	1.903453	-1.264152	2
19	哈尔滨	1.320269	1.248651	.2435345	-.9017014	-.2053833	2
20	南京	1.431901	1.47261	2.34706	2.582076	1.460131	2
21	呼和浩特	1.083554	1.368374	1.634534	-.9864622	2.529019	2
22	深圳	2.218884	2.305276	2.208282	1.20659	1.707282	2

图 18.27　按文化水平进行的聚类分析结果图 4

可以看到第 2 类所代表的人均旅游消费支出特点是无论文化水平如何花费支出总体上相对较高，第 1 类所代表的人均旅游消费支出特点是无论文化水平如何花费支出总体上相对较低。

图 18.28 展示的是设定聚类数为 3，然后使用“K 个平均数的聚类分析”方法进行分析的结果。在输入第 7 条 Stata 命令并且分别按键盘上的回车键进行确认后，可以看到系统产生了一个新的变量：聚类变量_clus_2（cluster name: _clus_2）。

```
. cluster kmeans zv2 zv3 zv4 zv5 zv6,k(3)
cluster name:  _clus_2
```

图 18.28　按文化水平进行的聚类分析结果图 5

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 18.29 所示的_clus_2 数据。

从图 18.29 中可以看到所有的观测样本被分为 3 类，其中南京、呼和浩特、深圳属于第 2 类，长沙、银川、太原、哈尔滨、长春属于第 3 类，其他城市属于第 1 类。

	V1	zv2	zv3	zv4	zv5	zv6	_clus_1	_clus_2
1	上海	-.5926007	-.0724776	-.8812586	-.416383	.3084838	1	1
2	长沙	.0365715	.3126196	-.3634083	.6744207	-1.067826	1	3
3	郑州	-.7863714	-.4485261	.0919059	-.4306886	-.5517095	1	1
4	北京	-.2826713	-.4716702	-.1698035	-.3831225	-.4655777	1	1
5	沈阳	-1.344788	-1.344186	-1.108916	-.7246693	-.3830442	1	1
6	杭州	-.4090491	-.6413357	.0636361	-.4335499	-.3234491	1	1
7	青岛	-.2441242	.0014792	-1.091568	.0569543	1.958031	1	1
8	武汉	-.5188696	-.7948178	-.0430179	-.1713992	-.3173771	1	1
9	天津	-.8395354	-.5461491	-.0166757	-.5204565	-.2903906	1	1
10	乌鲁木齐	.2165015	-.633331	-.743979	-.3926	-.0477374	1	1
11	石家庄	-.7272571	-.4955104	-.7180651	-.3924212	-.1698511	1	1
12	广州	-.6140734	-.6843176	-.1061965	-.5361927	-.3762976	1	1
13	大连	-1.432748	-1.424755	-1.336573	-1.110742	-.6767467	1	1
14	苏州	-.8307394	-1.111875	-.3049413	-.4995345	-.3182766	1	1
15	无锡	-.0279754	-.1909826	-.9138117	-.7421936	.0015128	1	1
16	银川	.452441	.458445	.1251014	1.556363	-.8609294	2	3
17	太原	.1206509	1.265009	.3052139	.662261	-.6457124	2	3
18	长春	1.770029	.4274701	.7789464	1.903453	-1.264152	2	3
19	哈尔滨	1.320269	1.248651	.2435345	-.9017014	-.2053833	2	3
20	南京	1.431901	1.47261	2.34706	2.582076	1.460131	2	2
21	呼和浩特	1.083554	1.368374	1.634534	-.9864622	2.529019	2	2
22	深圳	2.218884	2.305276	2.208282	1.20659	1.707282	2	2

图 18.29 按文化水平进行的聚类分析结果图 6

为观测 3 类样本的特征，可以对数据进行排序操作，在主界面的“Command”文本框中输入操作命令：

```
sort _clus_2
```

并按键盘上的回车键进行确认，然后选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 18.30 所示的整理后的数据。

	V1	zv2	zv3	zv4	zv5	zv6	_clus_1	_clus_2
1	杭州	-.4090491	-.6413357	.0636361	-.4335499	-.3234491	1	1
2	郑州	-.7863714	-.4485261	.0919059	-.4306886	-.5517095	1	1
3	沈阳	-1.344788	-1.344186	-1.108916	-.7246693	-.3830442	1	1
4	无锡	-.0279754	-.1909826	-.9138117	-.7421936	.0015128	1	1
5	苏州	-.8307394	-1.111875	-.3049413	-.4995345	-.3182766	1	1
6	大连	-1.432748	-1.424755	-1.336573	-1.110742	-.6767467	1	1
7	上海	-.5926007	-.0724776	-.8812586	-.416383	.3084838	1	1
8	武汉	-.5188696	-.7948178	-.0430179	-.1713992	-.3173771	1	1
9	天津	-.8395354	-.5461491	-.0166757	-.5204565	-.2903906	1	1
10	石家庄	-.7272571	-.4955104	-.7180651	-.3924212	-.1698511	1	1
11	广州	-.6140734	-.6843176	-.1061965	-.5361927	-.3762976	1	1
12	北京	-.2826713	-.4716702	-.1698035	-.3831225	-.4655777	1	1
13	青岛	-.2441242	.0014792	-1.091568	.0569543	1.958031	1	1
14	乌鲁木齐	.2165015	-.633331	-.743979	-.3926	-.0477374	1	1
15	南京	1.431901	1.47261	2.34706	2.582076	1.460131	2	2
16	呼和浩特	1.083554	1.368374	1.634534	-.9864622	2.529019	2	2
17	深圳	2.218884	2.305276	2.208282	1.20659	1.707282	2	2
18	长春	1.770029	.4274701	.7789464	1.903453	-1.264152	2	3
19	长沙	.0365715	.3126196	-.3634083	.6744207	-1.067826	1	3
20	银川	.452441	.458445	.1251014	1.556363	-.8609294	2	3
21	太原	.1206509	1.265009	.3052139	.662261	-.6457124	2	3
22	哈尔滨	1.320269	1.248651	.2435345	-.9017014	-.2053833	2	3

图 18.30 按文化水平进行的聚类分析结果图 7

从图 18.30 中可以看到第 2 类所代表的人均旅游消费支出特点是无论文化水平如何花费支出总体上相对最高，第 1 类所代表的人均旅游消费支出特点是无论文化水平如何花费支出总体上相对最低，第 3 类则表示中等水平。

图 18.31 展示的是设定聚类数为 4，然后使用“K 个平均数的聚类分析”方法进行分析的结果。在输入第 8 条 Stata 命令并且分别按键盘上的回车键进行确认后，可以看到系统产生了一个新的变量：聚类变量_clus_3（cluster name: _clus_3）。

```
. cluster kmeans zv2 zv3 zv4 zv5 zv6,k(4)
cluster name: _clus_3
```

图 18.31 按文化水平进行的聚类分析结果图 8

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 18.32 所示的_clus_3 数据。

	V1	zv2	zv3	zv4	zv5	zv6	_clus_1	_clus_2	_clus_3
1	杭州	-.4090491	-.6413357	.0636361	-.4335499	-.3234491	1	1	1
2	郑州	-.7863714	-.4485261	.0919059	-.4306886	-.5517095	1	1	1
3	沈阳	-1.344788	-1.344186	-1.108916	-.7246693	-.3830442	1	1	2
4	无锡	-.0279754	-.1909826	-.9138117	-.7421936	.0015128	1	1	1
5	苏州	-.8307394	-1.111875	-.3049413	-.4995345	-.3182766	1	1	1
6	大连	-1.432748	-1.424755	-1.336573	-1.110742	-.6767467	1	1	2
7	上海	-.5926007	-.0724776	-.8812586	-.416383	.3084838	1	1	1
8	武汉	-.5188696	-.7948178	-.0430179	-.1713992	-.3173771	1	1	1
9	天津	-.8395354	-.5461491	-.0166757	-.5204565	-.2903906	1	1	1
10	石家庄	-.7272571	-.4955104	-.7180651	-.3924212	-.1698511	1	1	1
11	广州	-.6140734	-.6843176	-.1061965	-.5361927	-.3762976	1	1	1
12	北京	-.2826713	-.4716702	-.1698035	-.3831225	-.4655777	1	1	1
13	青岛	-.2441242	.0014792	-1.091568	.0569543	1.958031	1	1	1
14	乌鲁木齐	.2165015	-.633331	-.743979	-.3926	-.0477374	1	1	1
15	南京	1.431901	1.47261	2.34706	2.582076	1.460131	2	2	3
16	呼和浩特	1.083554	1.368374	1.634534	-.9864622	2.529019	2	2	3
17	深圳	2.218884	2.305276	2.208282	1.20659	1.707282	2	2	3
18	长春	1.770029	.4274701	.7789464	1.903453	-1.264152	2	3	4
19	长沙	.0365715	.3126196	-.3634083	.6744207	-1.067826	1	3	4
20	银川	.452441	.458445	.1251014	1.556363	-.8609294	2	3	4
21	太原	.1206509	1.265009	.3052139	.662261	-.6457124	2	3	4
22	哈尔滨	1.320269	1.248651	.2435345	-.9017014	-.2053833	2	3	4

图 18.32 按文化水平进行的聚类分析结果图 9

从图 18.32 中可以看到所有的观测样本被分为 4 类。其中大连、沈阳属于第 2 类，南京、呼和浩特、深圳属于第 3 类，长沙、长春、哈尔滨、银川、太原属于第 4 类，其他城市属于第 1 类。从图 18.32 中很难看出各个类别的特征，可以对数据进行排序操作，在主界面的“Command”文本框中输入操作命令：

```
sort _clus_3
```

并按键盘上的回车键进行确认，然后选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 18.33 所示的整理后的数据。

	V1	zv2	zv3	zv4	zv5	zv6	_clus_1	_clus_2	_clus_3
1	杭州	-.4090491	-.6413357	.0636361	-.4335499	-.3234491	1	1	1
2	北京	-.2826713	-.4716702	-.1698035	-.3831225	-.4655777	1	1	1
3	青岛	-.2441242	.0014792	-1.091568	.0569543	1.958031	1	1	1
4	上海	-.5926007	-.0724776	-.8812586	-.416383	.3084838	1	1	1
5	郑州	-.7863714	-.4485261	.0919059	-.4306886	-.5517095	1	1	1
6	苏州	-.8307394	-1.111875	-.3049413	-.4995345	-.3182766	1	1	1
7	广州	-.6140734	-.6843176	-.1061965	-.5361927	-.3762976	1	1	1
8	乌鲁木齐	.2165015	-.633331	-.743979	-.3926	-.0477374	1	1	1
9	天津	-.8395354	-.5461491	-.0166757	-.5204565	-.2903906	1	1	1
10	武汉	-.5188696	-.7948178	-.0430179	-.1713992	-.3173771	1	1	1
11	无锡	-.0279754	-.1909826	-.9138117	-.7421936	.0015128	1	1	1
12	石家庄	-.7272571	-.4955104	-.7180651	-.3924212	-.1698511	1	1	1
13	沈阳	-1.344788	-1.344186	-1.108916	-.7246693	-.3830442	1	1	2
14	大连	-1.432748	-1.424755	-1.336573	-1.110742	-.6767467	1	1	2
15	南京	1.431901	1.47261	2.34706	2.582076	1.460131	2	2	3
16	呼和浩特	1.083554	1.368374	1.634534	-.9864622	2.529019	2	2	3
17	深圳	2.218884	2.305276	2.208282	1.20659	1.707282	2	2	3
18	哈尔滨	1.320269	1.248651	.2435345	-.9017014	-.2053833	2	3	4
19	长春	1.770029	.4274701	.7789464	1.903453	-1.264152	2	3	4
20	银川	.452441	.458445	.1251014	1.556363	-.8609294	2	3	4
21	长沙	.0365715	.3126196	-.3634083	.6744207	-1.067826	1	3	4
22	太原	.1206509	1.265009	.3052139	.662261	-.6457124	2	3	4

图 18.33 按文化水平进行的聚类分析结果图 10

从图 18.33 中可以看出，第 3 类所代表的人均旅游消费支出特点是：无论文化水平如何花费支出总体上相对最高，第 2 类所代表的人均旅游消费支出特点是：无论文化水平如何花费支出总体上相对最低，第 4 类所代表的人均旅游消费支出特点是：无论文化水平如何花费支出总体上相对较高，第 1 类所代表的人均旅游消费支出特点是无论文化水平如何花费支出总体上相对较低。

在前面的章节中也提到过，划分聚类分析的特点是需要事先制定拟分类的数量。究竟分成多少类是合理的，这是没有定论的。用户需要根据自己的研究和需要及数据的实际特点加入自己的判断。在上面的分析中，我们尝试着把样本分别分为 2、3、4 类进行了研究，可以看出把数据分成两类是过于粗糙的，而且两个类别所包含的样本数量的差别也是比较大的，而把数据分成 3 类是比较合适的。读者可以再把数据分成 5 类、6 类或者其他数量的类别进行研究，观察分类情况，找出自己认为最优的分类。

通过聚类分析得到的研究结论是：按文化水平进行聚类时，长沙、银川、太原、哈尔滨、长春等城市的城镇居民无论文化水平如何，其 2007 年人均旅游消费支出都处于全国中档水平上；南京、呼和浩特、深圳等城市的城镇居民无论文化水平如何，其 2007 年人均旅游消费支出都处于全国高档水平上；除以上城市之外的其他城市的城镇居民无论文化水平如何，其 2007 年人均旅游消费支出都处于全国低档水平上。

18.3.4 各城市国内旅游出游人均花费按旅游目的进行的聚类分析

	下载资源:\video\chap18\…
	下载资源:\sample\chap18\案例18.4.dta

表 18.7 是 2007 年我国 22 个城市城镇居民国内旅游出游人均花费按旅游目的进行分类的数据。

表 18.7 我国 2007 年城镇居民国内旅游出游人均花费情况统计（按旅游目的分组）（单位：元/人）

城市	观光游览	探亲访友	商务	公务会议	度假休闲
北京	1 272.1	805.4	2 302.0	16 29.3	653.6
天津	971.6	646.6	1 244.7	1 231.8	1 026.1
石家庄	989.9	352.3	3 058.0	1 364.0	1 124.8
太原	1 331.2	1 462.2	0	0	1 824.3
呼和浩特	2 436.0	1 298.2	3 135.0	814.0	1 848.6
沈阳	385.9	358.6	530.1	1 576.5	474.8
大连	350.5	351.5	1 958.2	1 246.9	151.4
长春	2 332.9	2 624.6	4 594.0	3 742.5	1 330.4
哈尔滨	1 623.5	1 898.3	3 032.4	3 670.7	1 986.4
上海	936.5	2 104.3	877.3	2 738.5	650.7
南京	2 381.0	1 671.7	2783.0	0	2 227.2
无锡	1 066.5	1 113.3	970.2	0	1 168.2
苏州	595.3	903.3	0	0	114.4

（续表）

城市	观光游览	探亲访友	商务	公务会议	度假休闲
杭州	1 359.3	467.6	869.0	1 619.6	452.2
青岛	1 485.5	804.8	2 254.7	892.5	902.0
郑州	966.3	468.7	0	660.0	330.0
武汉	1 098.9	500.5	2 568.7	1 365.1	1 185.0
长沙	1 864.1	1 006.2	1 859.1	0	606.5
广州	785.0	1 195.6	64.9	1 480.5	750.0
深圳	3 911.8	1 572.7	2 983.4	948.6	1 989.7
银川	1 598.3	1 033.2	5 011.6	1 815.2	1 483.5
乌鲁木齐	1 315.0	1 398.6	4 671.3	2 129.6	407.3

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有 6 个变量，分别为“城市”“观光游览”“探亲访友”“商务”“公务会议”“度假休闲”。将这 6 个变量分别定义为 V1~V6，然后录入相关数据。录入完成后数据如图 18.34 所示。

	V1	V2	V3	V4	V5	V6
1	北京	1272.1	805.4	2302	1629.3	653.6
2	天津	971.6	646.6	1244.7	1231.8	1026.1
3	石家庄	989.9	352.3	3058	1364	1124.8
4	太原	1331.2	1462.2	0	0	1824.3
5	呼和浩特	2436	1298.2	3135	814	1748.6
6	沈阳	385.9	358.6	530.1	1576.5	474.8
7	大连	350.5	351.5	1958.2	1246.9	151.4
8	长春	2332.9	2624.6	4594	3742.5	1330.4
9	哈尔滨	1623.5	1798.3	3032.4	3670.7	1986.4
10	上海	936.5	2104.3	877.3	2738.5	650.7
11	南京	2381	1671.7	2783	0	2227.2
12	无锡	1066.5	1113.3	970.2	0	1168.2
13	苏州	595.3	903.3	0	0	114.4
14	杭州	1359.3	467.6	869	1619.6	452.2
15	青岛	1485.5	804.8	2254.7	892.5	902
16	郑州	966.3	468.7	0	660	330
17	武汉	1098.9	500.5	2568.7	1365.1	1175
18	长沙	1864.1	1006.2	1759.1	0	606.5
19	广州	785	1195.6	64.9	1480.5	750
20	深圳	3911.8	1572.7	2983.4	948.6	1989.7
21	银川	1598.3	1033.2	5011.6	1815.2	1483.5
22	乌鲁木齐	1315	1398.6	4671.3	2129.6	407.3

图 18.34　案例 18.4 数据

聚类分析的分析步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中分别输入如下命令并按键盘上的回车键进行确认。

- egen zv2=std(V2)：本命令旨在对 V2 变量数据进行标准化处理，标准化处理方式是使变量的平均数为 0 而且标准差为 1。
- egen zv3=std(V3)：本命令旨在对 V3 变量数据进行标准化处理，标准化处理方式是使变量的平均数为 0 而且标准差为 1。
- egen zv4=std(V4)：本命令旨在对 V4 变量数据进行标准化处理，标准化处理方式是使变量的平均数为 0 而且标准差为 1。
- egen zv5=std(V5)：本命令旨在对 V5 变量数据进行标准化处理，标准化处理方式是使变量的平均数为 0 而且标准差为 1。

- egen zv6=std(V6)：本命令旨在对 V6 变量数据进行标准化处理，标准化处理方式是使变量的平均数为 0 而且标准差为 1。
- cluster kmeans zv2 zv3 zv4 zv5 zv6,k(2)：本命令旨在对 V2~V6 的标准化变量进行“K 个平均数的聚类分析”，设定的聚类数是 2。
- cluster kmeans zv2 zv3 zv4 zv5 zv6,k(3)：本命令旨在对 V2~V6 的标准化变量进行“K 个平均数的聚类分析”，设定的聚类数是 3。
- cluster kmeans zv2 zv3 zv4 zv5 zv6,k(4)：本命令旨在对 V2~V6 的标准化变量进行“K 个平均数的聚类分析”，设定的聚类数是 4。

03 设置完毕后，按键盘上的回车键，等待输出结果。

在 Stata 14.0 主界面的结果窗口可以看到如图 18.35~图 18.44 所示的分析结果。

1．数据标准化处理

在分析过程中前 5 条 Stata 命令旨在对数据进行标准化处理，选择的标准化处理方式是使变量的平均数为 0 而且标准差为 1。之所以这样做是因为我们进行聚类分析的变量都是以不可比的单位进行的测度，它们具有极为不同的方差，我们对数据进行标准化处理可以避免使结果受到具有最大方差变量的影响。在输入前 5 条 Stata 命令并且分别按键盘上的回车键进行确认后，选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 18.35 所示的变换后的数据。

	V5	V6	zv2	zv3	zv4	zv5	zv6
1	1629.3	653.6	-.1730331	-.4610598	.1750351	.2922291	-.5874082
2	1231.8	1026.1	-.5455368	-.7200493	-.5062146	-.0771064	-.0002078
3	1364	1124.8	-.5228518	-1.200028	.6621482	.0457266	.1553807
4	0	1824.3	-.099772	.6101255	-1.308212	-1.221628	1.258056
5	814	1748.6	1.269753	.3426553	.7117616	-.4653036	1.138724
6	1576.5	474.8	-1.271578	-1.189753	-.9666525	.2431702	-.8692644
7	1246.9	151.4	-1.31546	-1.201333	-.0464855	-.0630763	-1.379065
8	3742.5	1330.4	1.141948	2.505902	1.651838	2.2557	.4794838
9	3670.7	1986.4	.2625669	1.158276	.6456533	2.188988	1.513587
10	2738.5	650.7	-.5890472	1.657337	-.7429413	1.322838	-.5919797
11	0	2227.2	1.201574	.9518024	.4849576	-1.221628	1.893178
12	0	1168.2	-.4278975	.0410992	-.6830831	-1.221628	.2237953
13	0	114.4	-1.012003	-.3013932	-1.308212	-1.221628	-1.437391
14	1619.6	452.2	-.0649388	-1.011983	-.7482892	.2832163	-.9048905
15	892.5	902	.0915003	-.4620385	.1445583	-.3923657	-.1958362
16	660	330	-.5521067	-1.010189	-1.308212	-.6083921	-1.097524
17	1365.1	1175	-.387734	-.9583261	.3468778	.0467487	.2345147
18	0	606.5	.5608178	-.133572	-.1747715	-1.221628	-.6616555
19	1480.5	750	-.7768486	.1753234	-1.266395	.1539722	-.4354455
20	948.6	1989.7	3.099173	.7903417	.6140812	-.3402406	1.518789
21	1815.2	1483.5	.2313288	-.0895374	1.920911	.4649573	.720827
22	2129.6	407.3	-.1198537	.5063993	1.701645	.7570809	-.9756699

图 18.35　按旅游目的进行的聚类分析结果图 1

2．K 个平均数的聚类分析

图 18.36 展示的是设定聚类数为 2，然后使用“K 个平均数的聚类分析”方法进行分析的结果。在输入第 6 条 Stata 命令并且分别按键盘上的回车键进行确认后，可以看到系统产生了一个新的聚类变量。

```
. cluster kmeans zv2 zv3 zv4 zv5 zv6,k(2)
cluster name:  _clus_1
```

图 18.36　按旅游目的进行的聚类分析结果图 2

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 18.37 所示的聚类数据。

	V1	zv2	zv3	zv4	zv5	zv6	_clus_1
1	北京	-.1730331	-.4610598	.1750351	.2922291	-.5874082	1
2	天津	-.5455368	-.7200493	-.5062146	-.0771064	-.0002078	1
3	石家庄	-.5228518	-1.200028	.6621482	.0457266	.1553807	1
4	太原	-.099772	.6101255	-1.308212	-1.221628	1.258056	1
5	呼和浩特	1.269753	.3426553	.7117616	-.4653036	1.138724	2
6	沈阳	-1.271578	-1.189753	-.9666525	.2431702	-.8692644	1
7	大连	-1.31546	-1.201333	-.0464855	-.0630763	-1.379065	1
8	长春	1.141948	2.505902	1.651838	2.2557	.4794838	2
9	哈尔滨	.2625669	1.158276	.6456533	2.188988	1.513587	2
10	上海	-.5890472	1.657337	-.7429413	1.322838	-.5919797	1
11	南京	1.201574	.9518024	.4849576	-1.221628	1.893178	2
12	无锡	-.4278975	.0410992	-.6830831	-1.221628	.2237953	1
13	苏州	-1.012003	-.3013932	-1.308212	-1.221628	-1.437391	1
14	杭州	-.0649388	-1.011983	-.7482892	.2832163	-.9048905	1
15	青岛	.0915003	-.4620385	.1445583	-.3923657	-.1958362	1
16	郑州	-.5521067	-1.010189	-1.308212	-.6083921	-1.097524	1
17	武汉	-.387734	-.9583261	.3468778	.0467487	.2345147	1
18	长沙	.5608178	-.133572	-.1747715	-1.221628	-.6616555	1
19	广州	-.7768486	.1753234	-1.266395	.1539722	-.4354455	1
20	深圳	3.099173	.7903417	.6140812	-.3402406	1.518789	2
21	银川	.2313288	-.0895374	1.920911	.4649573	.720827	2
22	乌鲁木齐	-.1198537	.5063993	1.701645	.7570809	-.9756699	2

图 18.37　按旅游目的进行的聚类分析结果图 3

从图 18.37 中可以看到所有的观测样本被分为两类，其中太原、青岛、大连、武汉、杭州、石家庄、无锡、郑州、广州、上海、天津、北京、苏州、沈阳、长沙被分到第 1 类，南京、深圳、哈尔滨、银川、长春、呼和浩特、乌鲁木齐被分到第 2 类。

为观测两类样本的特征，可以对数据进行排序操作，在主界面的“Command”文本框中输入操作命令：

```
sort _clus_1
```

并按键盘上的回车键进行确认，然后选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 18.38 所示的整理后的数据。

可以看到第 2 类所代表的人均旅游消费支出特点是无论旅游目的如何花费支出总体上相对较高，第 1 类所代表的人均旅游消费支出特点是无论旅游目的如何花费支出总体上相对较低。

	V1	zv2	zv3	zv4	zv5	zv6	_clus_1
1	太原	-.099772	.6101255	-1.308212	-1.221628	1.258056	1
2	青岛	.0915003	-.4620385	.1445583	-.3923657	-.1958362	1
3	大连	-1.31546	-1.201333	-.0464855	-.0630763	-1.379065	1
4	武汉	-.387734	-.9583261	.3468778	.0467487	.2345147	1
5	杭州	-.0649388	-1.011983	-.7482892	.2832163	-.9048905	1
6	石家庄	-.5228518	-1.200028	.6621482	.0457266	.1553807	1
7	无锡	-.4278975	.0410992	-.6830831	-1.221628	.2237953	1
8	郑州	-.5521067	-1.010189	-1.308212	-.6083921	-1.097524	1
9	广州	-.7768486	.1753234	-1.266395	.1539722	-.4354455	1
10	上海	-.5890472	1.657337	-.7429413	1.322838	-.5919797	1
11	天津	-.5455368	-.7200493	-.5062146	-.0771064	-.0002078	1
12	北京	-.1730331	-.4610598	.1750351	.2922291	-.5874082	1
13	苏州	-1.012003	-.3013932	-1.308212	-1.221628	-1.437391	1
14	沈阳	-1.271578	-1.189753	-.9666525	.2431702	-.8692644	1
15	长沙	.5608178	-.133572	-.1747715	-1.221628	-.6616555	1
16	南京	1.201574	.9518024	.4849576	-1.221628	1.893178	2
17	深圳	3.099173	.7903417	.6140812	-.3402406	1.518789	2
18	哈尔滨	.2625669	1.158276	.6456533	2.188988	1.513587	2
19	银川	.2313288	-.0895374	1.920911	.4649573	.720827	2
20	长春	1.141948	2.505902	1.651838	2.2557	.4794838	2
21	呼和浩特	1.269753	.3426553	.7117616	-.4653036	1.138724	2
22	乌鲁木齐	-.1198537	.5063993	1.701645	.7570809	-.9756699	2

图 18.38　按旅游目的进行的聚类分析结果图 4

图 18.39 展示的是设定聚类数为 3，然后使用“K 个平均数的聚类分析”方法进行分析的结果。在输入第 7 条 Stata 命令并且分别按键盘上的回车键进行确认后，可以看到系统产生了一个新的变量：聚类变量_clus_2（cluster name: _clus_2）。

```
. cluster kmeans zv2 zv3 zv4 zv5 zv6,k(3)
cluster name: _clus_2
```

图 18.39 按旅游目的进行的聚类分析结果图 5

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 18.40 所示的_clus_2 数据。

	v1	zv2	zv3	zv4	zv5	zv6	_clus_1	_clus_2
1	太原	-.099772	.6101255	-1.308212	-1.221628	1.258056	1	1
2	青岛	.0915003	-.4620385	.1445583	-.3923657	-.1958362	1	1
3	大连	-1.31546	-1.201333	-.0464855	-.0630763	-1.379065	1	1
4	武汉	-.387734	-.9583261	.3468778	.0467487	.2345147	1	1
5	杭州	-.0649388	-1.011983	-.7482892	.2832163	-.9048905	1	1
6	石家庄	-.5228518	-1.200028	.6621482	.0457266	.1553807	1	1
7	无锡	-.4278975	.0410992	-.6830831	-1.221628	.2237953	1	1
8	郑州	-.5521067	-1.010189	-1.308212	-.6083921	-1.097524	1	1
9	广州	-.7768486	.1753234	-1.266395	.1539722	-.4354455	1	1
10	上海	-.5890472	1.657337	-.7429413	1.322838	-.5919797	1	2
11	天津	-.5455368	-.7200493	-.5062146	-.0771064	-.0002078	1	1
12	北京	-.1730331	-.4610598	.1750351	.2922291	-.5874082	1	1
13	苏州	-1.012003	-.3013932	-1.308212	-1.221628	-1.437391	1	1
14	沈阳	-1.271578	-1.189753	-.9666525	.2431702	-.8692644	1	1
15	长沙	.5608178	-.133572	-.1747715	-1.221628	-.6616555	1	1
16	南京	1.201574	.9518024	.4849576	-1.221628	1.893178	2	3
17	深圳	3.099173	.7903417	.6140812	-.3402406	1.518789	2	3
18	哈尔滨	.2625669	1.158276	.6456533	2.188988	1.513587	2	2
19	银川	.2313288	-.0895374	1.920911	.4649573	.720827	2	2
20	长春	1.141948	2.505902	1.651838	2.2557	.4794838	2	2
21	呼和浩特	1.269753	.3426553	.7117616	-.4653036	1.138724	2	3
22	乌鲁木齐	-.1198537	.5063993	1.701645	.7570809	-.9756699	2	2

图 18.40 按旅游目的进行的聚类分析结果图 6

从图 18.40 中可以看到所有的观测样本被分为 3 类，其中广州、苏州、北京、天津、大连、长沙、青岛、沈阳、石家庄、郑州、杭州、太原、武汉、无锡属于第 1 类，乌鲁木齐、长春、哈尔滨、上海、银川属于第 2 类，呼和浩特、南京、深圳属于第 3 类。

为观测 3 类样本的特征，可以对数据进行排序操作，在主界面的“Command”文本框中输入操作命令：

```
sort _clus_2
```

并按键盘上的回车键进行确认，然后选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 18.41 所示的整理后的数据。

	v1	zv2	zv3	zv4	zv5	zv6	_clus_1	_clus_2
1	广州	-.7768486	.1753234	-1.266395	.1539722	-.4354455	1	1
2	苏州	-1.012003	-.3013932	-1.308212	-1.221628	-1.437391	1	1
3	北京	-.1730331	-.4610598	.1750351	.2922291	-.5874082	1	1
4	天津	-.5455368	-.7200493	-.5062146	-.0771064	-.0002078	1	1
5	大连	-1.31546	-1.201333	-.0464855	-.0630763	-1.379065	1	1
6	长沙	.5608178	-.133572	-.1747715	-1.221628	-.6616555	1	1
7	青岛	.0915003	-.4620385	.1445583	-.3923657	-.1958362	1	1
8	沈阳	-1.271578	-1.189753	-.9666525	.2431702	-.8692644	1	1
9	石家庄	-.5228518	-1.200028	.6621482	.0457266	.1553807	1	1
10	郑州	-.5521067	-1.010189	-1.308212	-.6083921	-1.097524	1	1
11	杭州	-.0649388	-1.011983	-.7482892	.2832163	-.9048905	1	1
12	太原	-.099772	.6101255	-1.308212	-1.221628	1.258056	1	1
13	武汉	-.387734	-.9583261	.3468778	.0467487	.2345147	1	1
14	无锡	-.4278975	.0410992	-.6830831	-1.221628	.2237953	1	1
15	乌鲁木齐	-.1198537	.5063993	1.701645	.7570809	-.9756699	2	2
16	长春	1.141948	2.505902	1.651838	2.2557	.4794838	2	2
17	哈尔滨	.2625669	1.158276	.6456533	2.188988	1.513587	2	2
18	上海	-.5890472	1.657337	-.7429413	1.322838	-.5919797	1	2
19	银川	.2313288	-.0895374	1.920911	.4649573	.720827	2	2
20	呼和浩特	1.269753	.3426553	.7117616	-.4653036	1.138724	2	3
21	南京	1.201574	.9518024	.4849576	-1.221628	1.893178	2	3
22	深圳	3.099173	.7903417	.6140812	-.3402406	1.518789	2	3

图 18.41　按旅游目的进行的聚类分析结果图 7

从图 18.41 中可以看到第 1 类所代表的人均旅游消费支出特点是“观光游览”最低、“探亲访友”最低、“商务”最低、“公务会议”中等、“度假休闲”最低；第 2 类所代表的人均旅游消费支出特点是“观光游览”中等、“探亲访友”最高、“商务”最高、“公务会议”最高、“度假休闲”中等；第 3 类所代表的人均旅游消费支出特点是“观光游览”最高、“探亲访友”中等、“商务”中等、“公务会议”最低、“度假休闲”最高。

图 18.42 展示的是设定聚类数为 4，然后使用“K 个平均数的聚类分析”方法进行分析的结果。在输入第 8 条 Stata 命令并且分别按键盘上的回车键进行确认后，可以看到系统产生了一个新的变量：聚类变量_clus_3（cluster name: _clus_3）。

```
. cluster kmeans zv2 zv3 zv4 zv5 zv6,k(4)
cluster name:  _clus_3
```

图 18.42　按旅游目的进行的聚类分析结果图 8

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 18.43 所示的_clus_3 数据。

从图 18.43 中可以看到所有的观测样本被分为 4 类，其中乌鲁木齐、武汉、北京、石家庄、青岛、银川属于第 1 类，呼和浩特、长春、哈尔滨、深圳属于第 2 类，大连、苏州、上海、郑州、沈阳、天津、广州、杭州属于第 3 类，长沙、无锡、太原、南京属于第 4 类。从图 18.43 中很难看出各个类别的特征，可以对数据进行排序操作，在主界面的“Command”文本框中输入操作命令：

```
sort _clus_3
```

	V1	zv2	zv3	zv4	zv5	zv6	_clus_1	_clus_2	_clus_3
1	广州	-.7768486	.1753234	-1.266395	.1539722	-.4354455	1	1	3
2	苏州	-1.012003	-.3013932	-1.308212	-1.221628	-1.437391	1	1	3
3	北京	-.1730331	-.4610598	.1750351	.2922291	-.5874082	1	1	1
4	天津	-.5455368	-.7200493	-.5062146	-.0771064	-.0002078	1	1	3
5	大连	-1.31546	-1.201333	-.0464855	-.0630763	-1.379065	1	1	3
6	长沙	.5608178	-.133572	-.1747715	-1.221628	-.6616555	1	1	4
7	青岛	.0915003	-.4620385	.1445583	-.3923657	-.1958362	1	1	1
8	沈阳	-1.271578	-1.189753	-.9666525	.2431702	-.8692644	1	1	3
9	石家庄	-.5228518	-1.200028	.6621482	.0457266	.1553807	1	1	1
10	郑州	-.5521067	-1.010189	-1.308212	-.6083921	-1.097524	1	1	3
11	杭州	-.0649388	-1.011983	-.7482892	.2832163	-.9048905	1	1	3
12	太原	-.099772	.6101255	-1.308212	-1.221628	1.258056	1	1	4
13	武汉	-.387734	-.9583261	.3468778	.0467487	.2345147	1	1	1
14	无锡	-.4278975	.0410992	-.6830831	-1.221628	.2237953	1	1	4
15	乌鲁木齐	-.1198537	.5063993	1.701645	.7570809	-.9756699	2	2	1
16	长春	1.141948	2.505902	1.651838	2.2557	.4794838	2	2	2
17	哈尔滨	.2625669	1.158276	.6456533	2.188988	1.513587	2	2	2
18	上海	-.5890472	1.657337	-.7429413	1.322838	-.5919797	1	2	3
19	银川	.2313288	-.0895374	1.920911	.4649573	.720827	2	2	1
20	呼和浩特	1.269753	.3426553	.7117616	-.4653036	1.138724	2	3	2
21	南京	1.201574	.9518024	.4849576	-1.221628	1.893178	2	3	4
22	深圳	3.099173	.7903417	.6140812	-.3402406	1.518789	2	3	2

图 18.43　按旅游目的进行的聚类分析结果图 9

并按键盘上的回车键进行确认，然后选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 18.44 所示的整理后的数据。

	V1	zv2	zv3	zv4	zv5	zv6	_clus_1	_clus_2	_clus_3
1	乌鲁木齐	-.1198537	.5063993	1.701645	.7570809	-.9756699	2	2	1
2	武汉	-.387734	-.9583261	.3468778	.0467487	.2345147	1	1	1
3	北京	-.1730331	-.4610598	.1750351	.2922291	-.5874082	1	1	1
4	石家庄	-.5228518	-1.200028	.6621482	.0457266	.1553807	1	1	1
5	青岛	.0915003	-.4620385	.1445583	-.3923657	-.1958362	1	1	1
6	银川	.2313288	-.0895374	1.920911	.4649573	.720827	2	2	1
7	呼和浩特	1.269753	.3426553	.7117616	-.4653036	1.138724	2	3	2
8	长春	1.141948	2.505902	1.651838	2.2557	.4794838	2	2	2
9	哈尔滨	.2625669	1.158276	.6456533	2.188988	1.513587	2	2	2
10	深圳	3.099173	.7903417	.6140812	-.3402406	1.518789	2	3	2
11	大连	-1.31546	-1.201333	-.0464855	-.0630763	-1.379065	1	1	3
12	苏州	-1.012003	-.3013932	-1.308212	-1.221628	-1.437391	1	1	3
13	上海	-.5890472	1.657337	-.7429413	1.322838	-.5919797	1	2	3
14	郑州	-.5521067	-1.010189	-1.308212	-.6083921	-1.097524	1	1	3
15	沈阳	-1.271578	-1.189753	-.9666525	.2431702	-.8692644	1	1	3
16	天津	-.5455368	-.7200493	-.5062146	-.0771064	-.0002078	1	1	3
17	广州	-.7768486	.1753234	-1.266395	.1539722	-.4354455	1	1	3
18	杭州	-.0649388	-1.011983	-.7482892	.2832163	-.9048905	1	1	3
19	长沙	.5608178	-.133572	-.1747715	-1.221628	-.6616555	1	1	4
20	无锡	-.4278975	.0410992	-.6830831	-1.221628	.2237953	1	1	4
21	太原	-.099772	.6101255	-1.308212	-1.221628	1.258056	1	1	4
22	南京	1.201574	.9518024	.4849576	-1.221628	1.893178	2	3	4

图 18.44　按旅游目的进行的聚类分析结果图 10

从图 18.44 中可以看到第 1 类所代表的人均旅游消费支出特点是“观光游览”较低、“探亲访友”较低、“商务”最高、“公务会议”较高、“度假休闲”较低；第 2 类所代表的人均旅游消费支出特点是“观光游览” 最高、“探亲访友”最高、“商务”较高、“公务会议”最高、“度假休闲”最高；第 3 类所代表的人均旅游消费支出特点是“观光游览”最低、“探亲访友”最低、“商务”最低、“公务会议”较低、“度假休闲”最低；第 4 类所代表的人均旅游消费支出特点是“观光游览”较高、“探亲访友”较高、“商务”较低、“公务会议”最低、“度假休闲”较高。

在前面的章节中也提到过，划分聚类分析的特点是需要事先制定拟分类的数量。究竟分成多少类是合理的，这是没有定论的。用户需要根据自己的研究和需要及数据的实际特点加入自己的判断。在上面的分析中，我们尝试着把样本分别分为 2、3、4 类进行了研究，可以看出把数据分成两类是过于粗糙的，而且两个类别所包含的样本数量的差别也是比较大的，而把数据分成 3 类是比较合适的。读者可以再把数据分成 5 类、6 类或者其他数量的类别进行研究，观察分类情况，找出自己认为最优的分类。

通过聚类分析得到的研究结论是：乌鲁木齐、武汉、北京、石家庄、青岛、银川的人均旅游消费支出特点是“观光游览”较低、“探亲访友”较低、“商务”最高、“公务会议”较高、“度假休闲”较低；呼和浩特、长春、哈尔滨、深圳的人均旅游消费支出特点是“观光游览”最高、“探亲访友”最高、“商务”较高、“公务会议”最高、“度假休闲”最高；大连、苏州、上海、郑州、沈阳、天津、广州、杭州的人均旅游消费支出特点是“观光游览”最低、“探亲访友”最低、“商务”最低、“公务会议”较低、“度假休闲”最低；长沙、无锡、太原、南京的人均旅游消费支出特点是“观光游览”较高、“探亲访友”较高、“商务”较低、“公务会议”最低、“度假休闲”较高。

18.3.5 各风景区按其自身特点进行的聚类分析

	下载资源:\video\chap18\…
	下载资源:\sample\chap18\案例18.5.dta

表 18.8 是 2007 年我国部分国家级风景名胜区的统计数据。我们选取了 26 个著名的风景区并查找了相关资料，包括风景名胜区面积、游人量、景区资金收入、景区资金支出等，准备按照这些特征变量对景区名称进行聚类分析。

表 18.8 部分国家级风景名胜区数据统计

风景名胜区名称	风景名胜区面积/平方千米	游人量/万人次	景区资金收入/万元	景区资金支出/万元
十八重溪	62	7	205	210
青云山	52	61	3 700	7 500
鼓山	50	235	714	783
鼓浪屿	209	1 108	38 559	31 077
玉华洞	45	75	920	2 100
金湖	140	60	2 400	1 800
桃源洞	29	14	603	550
清源山	62	29	978	1 045
武夷山	79	250	25 000	29 000
冠豸山	123	135	985	1 403
鸳鸯溪	66	30	2 438	5 550
太姥山	320	115	47 800	3 500
梅岭	154	100	3 088	2 221
高岭	109	11	1 849	3 600
云居山	680	103	3 630	43 900
仙女湖	198	65	2 204	2 000
三百山	138	23	540	3051
武功山	445	6	620	900
井冈山	333	350	25 000	59 900
龟峰	39	48	1 180	1 180
三清山	229	103	9 281	75 000
青岛崂山	462	2 641	13 589	13 383
博山	73	370	4 100	3530

（续表）

风景名胜区名称	风景名胜区面积/平方千米	游人量/万人次	景区资金收入/万元	景区资金支出/万元
胶东半岛	92	669	36 785	27 641
青州	59	33	485	510
泰山	426	233	14 742	15 254

在用 Stata 进行分析之前，要把数据录入到 Stata 中。本例中有 5 个变量，分别为“风景名胜区名称”“风景名胜区面积”“游人量”“景区资金收入”“景区资金支出”。我们将这 5 个变量分别定义为 V1~V5，然后录入相关数据。录入完成后数据如图 18.45 所示。

	V1	V2	V3	V4	V5
1	十八重溪	62	7	205	210
2	青云山	52	61	3700	7500
3	鼓山	50	235	714	783
4	鼓浪屿	209	1108	38559	31077
5	玉华洞	45	75	920	2100
6	金湖	140	60	2400	1700
7	桃源洞	29	14	603	550
8	清源山	62	29	978	1045
9	武夷山	79	250	25000	29000
10	冠豸山	123	135	985	1403
11	鸳鸯溪	66	30	2438	5550
12	太姥山	320	115	47800	3500
13	梅岭	154	100	3088	2221
14	高岭	109	11	1849	3600
15	云居山	680	103	3630	43900
16	仙女湖	198	65	2204	2000
17	三百山	138	23	540	3051
18	武功山	445	6	620	900
19	井冈山	333	350	25000	59900
20	龟峰	39	48	1180	1180
21	三清山	229	103	9281	75000
22	青岛崂山	462	2641	13589	13383
23	博山	73	370	4100	3530
24	胶东半岛	92	669	36785	27641
25	青州	59	33	485	510
26	泰山	426	233	14742	15254

图 18.45　案例 18.5 数据

聚类分析的分析步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中分别输入如下命令并按键盘上的回车键进行确认。

- egen zv2=std(V2)：本命令旨在对 V2 变量数据进行标准化处理，标准化处理方式是使变量的平均数为 0 而且标准差为 1。
- egen zv3=std(V3)：本命令旨在对 V3 变量数据进行标准化处理，标准化处理方式是使变量的平均数为 0 而且标准差为 1。
- egen zv4=std(V4)：本命令旨在对 V4 变量数据进行标准化处理，标准化处理方式是使变量的平均数为 0 而且标准差为 1。
- egen zv5=std(V5)：本命令旨在对 V5 变量数据进行标准化处理，标准化处理方式是使变量的平均数为 0 而且标准差为 1。
- cluster kmeans zv2 zv3 zv4 zv5,k(2)：本命令旨在对 V2~V5 的标准化变量进行“K 个平均数的聚类分析”，设定的聚类数是 2。
- cluster kmeans zv2 zv3 zv4 zv5,k(3)：本命令旨在对 V2~V5 的标准化变量进行“K 个平均数的聚类分析”，设定的聚类数是 3。
- cluster kmeans zv2 zv3 zv4 zv5,k(4)：本命令旨在对 V2~V5 的标准化变量进行“K 个平均数的聚类分析”，设定的聚类数是 4。

03 设置完毕后，按键盘上的回车键，等待输出结果。

在 Stata 14.0 主界面的结果窗口可以看到如图 18.46~图 18.55 所示的分析结果。

1．数据标准化处理

在分析过程中前 4 条 Stata 命令旨在对数据进行标准化处理，选择的标准化处理方式是使变量的平均数为 0 而且标准差为 1。之所以这样做是因为我们进行聚类分析的变量都是以不可比的单位进行的测度，它们具有极为不同的方差，对数据进行标准化处理可以避免使结果受到具有最大方差变量的影响。在输入前 4 条 Stata 命令并且分别按键盘上的回车键进行确认后，选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 18.46 所示的变换后的数据。

	V1	V2	V3	V4	V5	zv2	zv3	zv4	zv5
1	十八重溪	62	7	205	210	-.7046929	-.4746873	-.6621238	-.6389999
2	青云山	52	61	3700	7500	-.7645297	-.3750966	-.4072483	-.2731214
3	鼓山	50	235	714	783	-.7764971	-.0541932	-.6250046	-.6102416
4	鼓浪屿	209	1108	38559	31077	.1749075	1.555856	2.134871	.9101871
5	玉华洞	45	75	920	2100	-.8064154	-.3492768	-.6099819	-.5441425
6	金湖	140	60	2400	1700	-.2379662	-.3769409	-.5020518	-.5642182
7	桃源洞	29	14	603	550	-.9021542	-.4617774	-.6330993	-.6219357
8	清源山	62	29	978	1045	-.7046929	-.4341133	-.6057522	-.597092
9	武夷山	79	250	25000	29000	-.6029704	-.0265291	1.14607	.8059443
10	冠豸山	123	135	985	1403	-.3396887	-.2386205	-.6052417	-.5791243
11	鸳鸯溪	66	30	2438	5550	-.6807582	-.432269	-.4992806	-.3709901
12	太姥山	320	115	47800	3500	.8390955	-.2755059	2.808778	-.4738778
13	梅岭	154	100	3088	2221	-.1541947	-.30317	-.4518788	-.5380697
14	高岭	109	11	1849	3600	-.4234602	-.4673102	-.5422338	-.4688589
15	云居山	680	103	3630	43900	2.993219	-.2976372	-.4123531	1.553762
16	仙女湖	198	65	2204	2000	.109087	-.3677195	-.5163452	-.5491614
17	三百山	138	23	540	3051	-.2499335	-.445179	-.6376936	-.4964127
18	武功山	445	6	620	900	1.587055	-.4765316	-.6318596	-.6043695
19	井冈山	333	350	25000	59900	.9168833	.1578981	1.14607	2.356788
20	龟峰	39	48	1180	1180	-.8423175	-.3990721	-.5910212	-.5903165
21	三清山	229	103	9281	75000	.294581	-.2976372	-.0002496	3.114643
22	青岛崂山	462	2641	13589	13383	1.688777	4.383126	.3139145	.0221411
23	博山	73	370	4100	3530	-.6388725	.1947835	-.378078	-.4723721
24	胶东半岛	92	669	36785	27641	-.5251827	.7462209	2.0055	.7377374
25	青州	59	33	485	510	-.722644	-.4267362	-.6417046	-.6239432
26	泰山	426	233	14742	15254	1.473365	-.0578818	.3979979	.1160449

图 18.46　各风景区按其自身特点进行的聚类分析结果图 1

2．K 个平均数的聚类分析

（1）聚类数为 2

图 18.47 展示的是设定聚类数为 2，然后使用“K 个平均数的聚类分析”方法进行分析的结果。在输入第 5 条 Stata 命令并且分别按键盘上的回车键进行确认后，可以看到系统产生了一个新的聚类变量。

```
. cluster kmeans zv2 zv3 zv4 zv5,k(2)
cluster name: _clus_1
```

图 18.47　各风景区按其自身特点进行的聚类分析结果图 2

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 18.48 所示的聚类数据。

	V1	zv2	zv3	zv4	zv5	_clus_1
1	十八重溪	-.7046929	-.4746873	-.6621238	-.6389999	2
2	青云山	-.7645297	-.3750966	-.4072483	-.2731214	2
3	鼓山	-.7764971	-.0541932	-.6250046	-.6102416	2
4	鼓浪屿	.1749075	1.555856	2.134871	.9101871	1
5	玉华洞	-.8064154	-.3492768	-.6099819	-.5441425	2
6	金湖	-.2379662	-.3769409	-.5020518	-.5642182	2
7	桃源洞	-.9021542	-.4617774	-.6330993	-.6219357	2
8	清源山	-.7046929	-.4341133	-.6057522	-.597092	2
9	武夷山	-.6029704	-.0265291	1.14607	.8059443	1
10	冠豸山	-.3396887	-.2386205	-.6052417	-.5791243	2
11	鸳鸯溪	-.6807582	-.432269	-.4992806	-.3709901	2
12	太姥山	.8390955	-.2755059	2.808778	-.4738778	1
13	梅岭	-.1541947	-.30317	-.4518788	-.5380697	2
14	高岭	-.4234602	-.4673102	-.5422338	-.4688589	2
15	云居山	2.993219	-.2976372	-.4123531	1.553762	1
16	仙女湖	.109087	-.3677195	-.5163452	-.5491614	2
17	三百山	-.2499335	-.445179	-.6376936	-.4964127	2
18	武功山	1.587055	-.4765316	-.6318596	-.6043695	2
19	井冈山	.9168833	.1578981	1.14607	2.356788	1
20	龟峰	-.8423175	-.3990721	-.5910212	-.5903165	2
21	三清山	.294581	-.2976372	-.0002496	3.114643	1
22	青岛崂山	1.688777	4.383126	.3139145	.0221411	1
23	博山	-.6388725	.1947835	-.378078	-.4723721	2
24	胶东半岛	-.5251827	.7462209	2.0055	.7377374	1
25	青州	-.722644	-.4267362	-.6417046	-.6239432	2
26	泰山	1.473365	-.0578818	.3979979	.1160449	1

图 18.48　各风景区按其自身特点进行的聚类分析结果图 3

从图 18.48 中可以看到所有的观测样本被分为两类，其中武夷山、三清山、太姥山、胶东半岛、鼓浪屿、云居山、泰山、青岛崂山、井冈山被分到第 1 类，梅岭、清源山、青州、桃源洞、武功山、仙女湖、高岭、鸳鸯溪、金湖、冠豸山、三百山、龟峰、博山、鼓山、青云山、十八重溪、玉华洞被分到第 2 类。

为观测两类样本的特征，可以对数据进行排序操作，在主界面的“Command”文本框中输入操作命令：

```
sort _clus_1
```

并按键盘上的回车键进行确认，然后选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 18.49 所示的整理后的数据。

可以看到第 1 类所代表的景区特点是“风景名胜区面积”“游人量”“景区资金收入”“景区资金支出”都相对较高，第 2 类所代表的景区特点是“风景名胜区面积”“游人量”“景区资金收入”“景区资金支出”都相对较低。

	V1	zv2	zv3	zv4	zv5	_clus_1
1	武夷山	-.6029704	-.0265291	1.14607	.8059443	1
2	三清山	.294581	-.2976372	-.0002496	3.114643	1
3	太姥山	.8390955	-.2755059	2.808778	-.4738778	1
4	胶东半岛	-.5251827	.7462209	2.0055	.7377374	1
5	鼓浪屿	.1749075	1.555856	2.134871	.9101871	1
6	云居山	2.993219	-.2976372	-.4123531	1.553762	1
7	泰山	1.473365	-.0578818	.3979979	.1160449	1
8	青岛崂山	1.688777	4.383126	.3139145	.0221411	1
9	井冈山	.9168833	.1578981	1.14607	2.356788	1
10	梅岭	-.1541947	-.30317	-.4518788	-.5380697	2
11	清源山	-.7046929	-.4341133	-.6057522	-.597092	2
12	青州	-.722644	-.4267362	-.6417046	-.6239432	2
13	桃源洞	-.9021542	-.4617774	-.6330993	-.6219357	2
14	武功山	1.587055	-.4765316	-.6318596	-.6043695	2
15	仙女湖	.109087	-.3677195	-.5163452	-.5491614	2
16	高岭	-.4234602	-.4673102	-.5422338	-.4688589	2
17	鸳鸯溪	-.6807582	-.432269	-.4992806	-.3709901	2
18	金湖	-.2379662	-.3769409	-.5020518	-.5642182	2
19	冠豸山	-.3396887	-.2386205	-.6052417	-.5791243	2
20	三百山	-.2499335	-.445179	-.6376936	-.4964127	2
21	龟峰	-.8423175	-.3990721	-.5910212	-.5903165	2
22	博山	-.6388725	.1947835	-.378078	-.4723721	2
23	鼓山	-.7764971	-.0541932	-.6250046	-.6102416	2
24	青云山	-.7645297	-.3750966	-.4072483	-.2731214	2
25	十八重溪	-.7046929	-.4746873	-.6621238	-.6389999	2
26	玉华洞	-.8064154	-.3492768	-.6099819	-.5441425	2

图 18.49　各风景区按其自身特点进行的聚类分析结果图 4

（2）聚类数为 3

图 18.50 展示的是设定聚类数为 3，然后使用“K 个平均数的聚类分析”方法进行分析的结果。在输入第 6 条 Stata 命令并且分别按键盘上的回车键进行确认后，可以看到系统产生了一个新的变量：聚类变量_clus_2（cluster name: _clus_2）。

```
. cluster kmeans zv2 zv3 zv4 zv5,k(3)
cluster name: _clus_2
```

图 18.50　各风景区按其自身特点进行的聚类分析结果图 5

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 18.51 所示的_clus_2 数据。

	V1	zv2	zv3	zv4	zv5	_clus_1	_clus_2
1	武夷山	-.6029704	-.0265291	1.14607	.8059443	1	1
2	三清山	.294581	-.2976372	-.0002496	3.114643	1	1
3	太姥山	.8390955	-.2755059	2.808778	-.4738778	1	1
4	胶东半岛	-.5251827	.7462209	2.0055	.7377374	1	1
5	鼓浪屿	.1749075	1.555856	2.134871	.9101871	1	1
6	云居山	2.993219	-.2976372	-.4123531	1.553762	1	1
7	泰山	1.473365	-.0578818	.3979979	.1160449	1	1
8	青岛崂山	1.688777	4.383126	.3139145	.0221411	1	3
9	井冈山	.9168833	.1578981	1.14607	2.356788	1	1
10	梅岭	-.1541947	-.30317	-.4518788	-.5380697	2	2
11	清源山	-.7046929	-.4341133	-.6057522	-.597092	2	2
12	青州	-.722644	-.4267362	-.6417046	-.6239432	2	2
13	桃源洞	-.9021542	-.4617774	-.6330993	-.6219357	2	2
14	武功山	1.587055	-.4765316	-.6318596	-.6043695	2	2
15	仙女湖	.109087	-.3677195	-.5163452	-.5491614	2	2
16	高岭	-.4234602	-.4673102	-.5422338	-.4688589	2	2
17	鸳鸯溪	-.6807582	-.432269	-.4992806	-.3709901	2	2
18	金湖	-.2379662	-.3769409	-.5020518	-.5642182	2	2
19	冠豸山	-.3396887	-.2386205	-.6052417	-.5791243	2	2
20	三百山	-.2499335	-.445179	-.6376936	-.4964127	2	2
21	龟峰	-.8423175	-.3990721	-.5910212	-.5903165	2	2
22	博山	-.6388725	.1947835	-.378078	-.4723721	2	2
23	鼓山	-.7764971	-.0541932	-.6250046	-.6102416	2	2
24	青云山	-.7645297	-.3750966	-.4072483	-.2731214	2	2
25	十八重溪	-.7046929	-.4746873	-.6621238	-.6389999	2	2
26	玉华洞	-.8064154	-.3492768	-.6099819	-.5441425	2	2

图 18.51　各风景区按其自身特点进行的聚类分析结果图 6

从图 18.51 中可以看到所有的观测样本被分为 3 类，其中武夷山、三清山、胶东半岛、云居山、泰山、太姥山、鼓浪屿、井冈山属于第 1 类，十八重溪、金湖、鼓山、武功山、清源山、冠豸山、高岭、鸳鸯溪、仙女湖、博山、桃源洞、梅岭、玉华洞、青云山、青州、三百山、龟峰属于第 2 类，青岛崂山属于第 3 类。

为观测 3 类样本的特征，可以对数据进行排序操作，在主界面的“Command”文本框中输入操作命令：

```
sort _clus_2
```

并按键盘上的回车键进行确认，然后选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 18.52 所示的整理后的数据。

	V1	zv2	zv3	zv4	zv5	_clus_1	_clus_2
1	武夷山	-.6029704	-.0265291	1.14607	.8059443	1	1
2	三清山	.294581	-.2976372	-.0002496	3.114643	1	1
3	胶东半岛	-.5251827	.7462209	2.0055	.7377374	1	1
4	云居山	2.993219	-.2976372	-.4123531	1.553762	1	1
5	泰山	1.473365	-.0578818	.3979979	.1160449	1	1
6	太姥山	.8390955	-.2755059	2.808778	-.4738778	1	1
7	鼓浪屿	.1749075	1.555856	2.134871	.9101871	1	1
8	井冈山	.9168833	.1578981	1.14607	2.356788	1	1
9	十八重溪	-.7046929	-.4746873	-.6621238	-.6389999	2	2
10	金湖	-.2379662	-.3769409	-.5020518	-.5642182	2	2
11	鼓山	-.7764971	-.0541932	-.6250046	-.6102416	2	2
12	武功山	1.587055	-.4765316	-.6318596	-.6043695	2	2
13	清源山	-.7046929	-.4341133	-.6057522	-.597092	2	2
14	冠豸山	-.3396887	-.2386205	-.6052417	-.5791243	2	2
15	高岭	-.4234602	-.4673102	-.5422338	-.4688589	2	2
16	鸳鸯溪	-.6807582	-.432269	-.4992806	-.3709901	2	2
17	仙女湖	.109087	-.3677195	-.5163452	-.5491614	2	2
18	博山	-.6388725	.1947835	-.378078	-.4723721	2	2
19	桃源洞	-.9021542	-.4617774	-.6330993	-.6219357	2	2
20	梅岭	-.1541947	-.30317	-.4518788	-.5380697	2	2
21	玉华洞	-.8064154	-.3492768	-.6099819	-.5441425	2	2
22	青云山	-.7645297	-.3750966	-.4072483	-.2731214	2	2
23	青州	-.722644	-.4267362	-.6417046	-.6239432	2	2
24	三百山	-.2499335	-.445179	-.6376936	-.4964127	2	2
25	龟峰	-.8423175	-.3990721	-.5910212	-.5903165	2	2
26	青岛崂山	1.688777	4.383126	.3139145	.0221411	1	3

图 18.52　各风景区按其自身特点进行的聚类分析结果图 7

从图 18.52 中可以看到第 1 类所代表的景区特点是“风景名胜区面积”“游人量”“景区资金收入”“景区资金支出”都相对处于中等水平；第 2 类所代表的景区特点是“风景名胜区面积”“游人量”“景区资金收入”“景区资金支出”都相对处于低等水平；第 3 类所代表的景区特点是“风景名胜区面积”“游人量”“景区资金收入”“景区资金支出”都相对处于高等水平。

（3）聚类数为 4

图 18.53 展示的是设定聚类数为 4，然后使用“K 个平均数的聚类分析”方法进行分析的结果。在输入第 7 条 Stata 命令并且分别按键盘上的回车键进行确认后，可以看到系统产生了一个新的变量：聚类变量_clus_3（cluster name: _clus_3）。

```
. cluster kmeans zv2 zv3 zv4 zv5,k(4)
cluster name: _clus_3
```

图 18.53　各风景区按其自身特点进行的聚类分析结果图 8

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 18.54 所示的_clus_3 数据。

	v1	zv2	zv3	zv4	zv5	_clus_1	_clus_2	_clus_3
1	武夷山	-.6029704	-.0265291	1.14607	.8059443	1	1	4
2	三清山	.294581	-.2976372	-.0002496	3.114643	1	1	1
3	胶东半岛	-.5251827	.7462209	2.0055	.7377374	1	1	4
4	云居山	2.993219	-.2976372	-.4123531	1.553762	1	1	1
5	泰山	1.473365	-.0578818	.3979979	.1160449	1	1	2
6	太姥山	.8390955	-.2755059	2.808778	-.4738778	1	1	4
7	鼓浪屿	.1749075	1.555856	2.134871	.9101871	1	1	4
8	井冈山	.9168833	.1578981	1.14607	2.356788	1	1	4
9	十八重溪	-.7046929	-.4746873	-.6621238	-.6389999	2	2	3
10	金湖	-.2379662	-.3769409	-.5020518	-.5642182	2	2	3
11	鼓山	-.7764971	-.0541932	-.6250046	-.6102416	2	2	3
12	武功山	1.587055	-.4765316	-.6318596	-.6043695	2	2	2
13	清源山	-.7046929	-.4341133	-.6057522	-.597092	2	2	3
14	冠豸山	-.3396887	-.2386205	-.6052417	-.5791243	2	2	3
15	高岭	-.4234602	-.4673102	-.5422338	-.4688589	2	2	3
16	鸳鸯溪	-.6807582	-.432269	-.4992806	-.3709901	2	2	3
17	仙女湖	.109087	-.3677195	-.5163452	-.5491614	2	2	3
18	博山	-.6388725	.1947835	-.378078	-.4723721	2	2	3
19	桃源洞	-.9021542	-.4617774	-.6330993	-.6219357	2	2	3
20	梅岭	-.1541947	-.30317	-.4518788	-.5380697	2	2	3
21	玉华洞	-.8064154	-.3492768	-.6099819	-.5441425	2	2	3
22	青云山	-.7645297	-.3750966	-.4072483	-.2731214	2	2	3
23	青州	-.722644	-.4267362	-.6417046	-.6239432	2	2	3
24	三百山	-.2499335	-.445179	-.6376936	-.4964127	2	2	3
25	龟峰	-.8423175	-.3990721	-.5910212	-.5903165	2	2	3
26	青岛崂山	1.688777	4.383126	.3139145	.0221411	1	3	1

图 18.54　各风景区按其自身特点进行的聚类分析结果图 9

从图 18.54 中可以看到所有的观测样本被分为 4 类，其中三清山、云居山、青岛崂山属于第 1 类，武功山、泰山属于第 2 类，龟峰、鸳鸯溪、梅岭、博山、桃源洞、十八重溪、金湖、仙女湖、玉华洞、青州、清源山、青云山、三百山、冠豸山、鼓山、高岭属于第 3 类，鼓浪屿、太姥山、井冈山、武夷山、胶东半岛属于第 4 类。从图 18.54 中很难看出各个类别的特征，可以对数据进行排序操作，在主界面的“Command”文本框中输入操作命令：

```
sort _clus_3
```

并按键盘上的回车键进行确认，然后选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 18.55 所示的整理后的数据。

	v1	zv2	zv3	zv4	zv5	_clus_1	_clus_2	_clus_3
1	三清山	.294581	-.2976372	-.0002496	3.114643	1	1	1
2	云居山	2.993219	-.2976372	-.4123531	1.553762	1	1	1
3	青岛崂山	1.688777	4.383126	.3139145	.0221411	1	3	1
4	武功山	1.587055	-.4765316	-.6318596	-.6043695	2	2	2
5	泰山	1.473365	-.0578818	.3979979	.1160449	1	1	2
6	龟峰	-.8423175	-.3990721	-.5910212	-.5903165	2	2	3
7	鸳鸯溪	-.6807582	-.432269	-.4992806	-.3709901	2	2	3
8	梅岭	-.1541947	-.30317	-.4518788	-.5380697	2	2	3
9	博山	-.6388725	.1947835	-.378078	-.4723721	2	2	3
10	桃源洞	-.9021542	-.4617774	-.6330993	-.6219357	2	2	3
11	十八重溪	-.7046929	-.4746873	-.6621238	-.6389999	2	2	3
12	金湖	-.2379662	-.3769409	-.5020518	-.5642182	2	2	3
13	仙女湖	.109087	-.3677195	-.5163452	-.5491614	2	2	3
14	玉华洞	-.8064154	-.3492768	-.6099819	-.5441425	2	2	3
15	青州	-.722644	-.4267362	-.6417046	-.6239432	2	2	3
16	清源山	-.7046929	-.4341133	-.6057522	-.597092	2	2	3
17	青云山	-.7645297	-.3750966	-.4072483	-.2731214	2	2	3
18	三百山	-.2499335	-.445179	-.6376936	-.4964127	2	2	3
19	冠豸山	-.3396887	-.2386205	-.6052417	-.5791243	2	2	3
20	鼓山	-.7764971	-.0541932	-.6250046	-.6102416	2	2	3
21	高岭	-.4234602	-.4673102	-.5422338	-.4688589	2	2	3
22	鼓浪屿	.1749075	1.555856	2.134871	.9101871	1	1	4
23	太姥山	.8390955	-.2755059	2.808778	-.4738778	1	1	4
24	井冈山	.9168833	.1578981	1.14607	2.356788	1	1	4
25	武夷山	-.6029704	-.0265291	1.14607	.8059443	1	1	4
26	胶东半岛	-.5251827	.7462209	2.0055	.7377374	1	1	4

图 18.55　各风景区按其自身特点进行的聚类分析结果图 10

从图 18.55 中可以看出，第 1 类所代表的景区特点是“风景名胜区面积”非常大、“游人量”比较大、“景区资金收入”比较大、“景区资金支出”非常大；第 2 类所代表的景区特点是“风景名胜区面积”比较大、“游人量”比较小、“景区资金收入”比较小、“景区资金支出”比较小；第 3 类所代表的景区特点是“风景名胜区面积”非常小、“游人量”非常小、“景区资金收入”非常小、“景区资金支出”非常小；第 4 类所代表的景区特点是“风景名胜区面积”比较小、“游人量”非常大、“景区资金收入”非常大、“景区资金支出”比较大。

在前面的章节中也提到过，划分聚类分析的特点是需要事先制定拟分类的数量。究竟分成多少类是合理的，这是没有定论的。用户需要根据自己的研究和需要及数据的实际特点加入自己的判断。在上面的分析中，我们尝试着把样本分别分为 2、3、4 类进行了研究，可以看出把数据分成两类是过于粗糙的，而且两个类别所包含的样本数量的差别也是比较大的，而把数据分成 3 类是比较合适的。读者可以再把数据分成 5 类、6 类或者其他数量的类别进行研究，观察分类情况，找出自己认为最优的分类。

通过聚类分析得到的研究结论是：三清山、云居山、青岛崂山的“风景名胜区面积”“游人量”“景区资金收入”“景区资金支出”都相对处于中等水平；武功山、泰山的“风景名胜区面积”“游人量”“景区资金收入”“景区资金支出”都相对处于低等水平；龟峰、鸳鸯溪、梅岭、博山、桃源洞、十八重溪、金湖、仙女湖、玉华洞、青州、清源山、青云山、三百山、冠豸山、鼓山、高岭的“风景名胜区面积”“游人量”“景区资金收入”“景区资金支出”都相对处于高等水平。

18.4　研究结论

根据以上所做的分析，可以比较有把握地得出以下结论：

- 按性别和年龄进行分类，青岛、长沙、银川、乌鲁木齐、太原、哈尔滨等城市的城镇居民无论男女老少，其 2007 年人均旅游消费支出都处于全国中档水平上；长春、南京、呼和浩特、深圳等城市的城镇居民无论男女老少，其 2007 年人均旅游消费支出都处于全国高档水平上；除以上城市之外的其他城市的城镇居民无论男女老少，其 2007 年人均旅游消费支出都处于全国低档水平上。
- 按职业进行分类，上海、郑州、北京、杭州、武汉、青岛、长沙、银川、乌鲁木齐、太原、哈尔滨、无锡等城市的城镇居民无论职业类型如何，其 2007 年人均旅游消费支出都处于全国中档水平上；长春、南京、呼和浩特、深圳等城市的城镇居民无论职业类型如何，其 2007 年人均旅游消费支出都处于全国高档水平上；除以上城市之外的其他城市的城镇居民无论职业类型如何，其 2007 年人均旅游消费支出都处于全国低档水平上。
- 按文化水平进行分类，长沙、银川、太原、哈尔滨、长春等城市的城镇居民无论文化水平如何，其 2007 年人均旅游消费支出都处于全国中档水平上；南京、呼和浩特、深圳等城市的城镇居民无论文化水平如何，其 2007 年人均旅游消费支出都处于全国高档水平上；除以上城市之外的其他城市的城镇居民无论文化水平如何，其 2007 年人均旅游消费支出都处于全国低档水平上。

- 按旅游目的进行分类，乌鲁木齐、武汉、北京、石家庄、青岛、银川的人均旅游消费支出特点是“观光游览”较低、“探亲访友”较低、“商务”最高、“公务会议”较高、“度假休闲”较低；呼和浩特、长春、哈尔滨、深圳的人均旅游消费支出特点是“观光游览”最高、“探亲访友”最高、“商务”较高、“公务会议”最高、“度假休闲”最高；大连、苏州、上海、郑州、沈阳、天津、广州、杭州的人均旅游消费支出特点是“观光游览”最低、“探亲访友”最低、“商务”最低、“公务会议”较低、“度假休闲”最低；长沙、无锡、太原、南京的人均旅游消费支出特点是“观光游览”较高、“探亲访友”较高、“商务”较低、“公务会议”最低、“度假休闲”较高。
- 三清山、云居山、青岛崂山的“风景名胜区面积”“游人量”“景区资金收入”“景区资金支出”都相对处于中等水平；武功山、泰山的“风景名胜区面积”“游人量”“景区资金收入”“景区资金支出”都相对处于低等水平；龟峰、鸳鸯溪、梅岭、博山、桃源洞、十八重溪、金湖、仙女湖、玉华洞、青州、清源山、青云山、三百山、冠豸山、鼓山、高岭的“风景名胜区面积”“游人量”“景区资金收入”“景区资金支出”都相对处于高等水平。

18.5 本章习题

（1）表 18.9 是 2006 年我国 22 个城市城镇居民国内旅游出游人均花费按性别和年龄进行分类的数据。试据此用本章介绍的方法将各城市按性别和年龄进行聚类。

表 18.9 我国 2006 年城镇居民国内旅游出游人均花费情况统计（按性别和年龄分组）（单位：元/人）

城市	男	女	65岁及以上	45~65岁	25~44岁	15~24岁	0~14岁
北京	939.5	796.3	591.3	874.3	1046.6	704.2	494.1
天津	808.9	716.2	821.7	843.5	779.5	493.5	468.2
石家庄	647.0	665.8	238.6	551.9	886.9	530.1	608.6
太原	1159.7	1857.1	1215.1	841.8	1569.7	3292.9	656.9
呼和浩特	2058.3	1800.2	1547.9	2233.1	2018.3	80.7	1116.0
沈阳	427.1	366.9	418.6	390.5	390.3	564.3	186.0
大连	309.8	249.4	180.5	237.8	403.3	277.5	113.0
长春	2553.8	1877.8	1161.3	3123.5	2062.4	1410.2	1101.9
哈尔滨	1408.2	1093.3	746.9	1349.1	1366.6	1450.0	331.6
上海	819.7	1116.1	2924.0	626.6	1019.0	1107.7	588.4
南京	1570.0	1138.6	2082.8	1185.7	1338.3	1327.9	994.0
无锡	1451.5	909.9	1215.9	635.7	1577.0	961.7	1350.4
苏州	3002.7	822.7	547.4	732.7	7429.0	686.4	2068.0
杭州	802.6	821.7	599.8	897.2	822.2	893.9	413.1
青岛	1347.8	1498.6	1336.0	1471.4	1583.6	1161.6	207.1
郑州	847.1	796.8	1077.3	825.5	854.3	596.5	660.7
武汉	1312.3	989.6	1166.3	1209.3	1134.8	884.5	527.3
长沙	1623.4	1115.4	1092.6	1495.3	1441.2	1209.4	446.4
广州	591.2	668.3	466.4	685.9	690.9	628.3	418.8
深圳	1867.1	1820.6	425.7	1903.1	1876.1	1865.9	1091.8
银川	1202.6	1414.0	961.7	1186.8	1370.3	1004.4	1608.2
乌鲁木齐	598.4	1019.6	528.0	763.3	1012.8	671.9	312.3

（2）表 18.10 是 2006 年我国 22 个城市城镇居民国内旅游出游人均花费按职业进行分类

的数据。试据此用本章介绍的方法将各城市按职业进行聚类。

表 18.10　我国 2006 年城镇居民国内旅游出游人均花费情况统计（按职业分组）（单位：元/人）

城市	公务员	企事业管理人员	技术人员	商贸人员	工人
北京	1622.3	1319.2	1090.0	647.6	718.6
天津	1529.9	942.0	777.1	613.8	599.4
石家庄	1226.2	836.7	938.9	636.4	481.7
太原	1045.0	1516.2	691.7	2113.1	1659.2
呼和浩特	1688.7	2256.5	2855.2	1265.4	561.0
沈阳	519.8	527.4	449.4	405.5	456.5
大连	637.3	354.7	552.4	387.2	239.5
长春	1870.1	2635.9	2640.7	1606.5	2056.1
哈尔滨	2746.9	2219.1	1351.4	1094.0	848.7
上海	1264.7	1116.1	1013.2	851.0	808.7
南京	2110.2	1201.5	1694.0	820.0	674.3
无锡	2108.2	1258.6	1829.0	818.0	397.7
苏州	9218.0	15195.7	1072.3	550.0	700.1
杭州	1325.1	1481.6	594.7	999.2	900.6
青岛	2115.9	2043.6	1279.4	2274.2	787.9
郑州	902.8	1020.2	961.8	183.0	1032.4
武汉	2344.5	1415.0	2133.4	840.0	915.2
长沙	1611.0	2181.8	2090.2	1136.9	402.7
广州	740.4	552.1	800.7	750.6	779.7
深圳	1834.6	1818.9	1851.2	2041.3	2549.6
银川	1554.7	1675.1	2129.3	1165.8	1016.1
乌鲁木齐	1892.2	831.1	1031.5	808.3	382.3

（3）表 18.11 是 2006 年我国 22 个城市城镇居民国内旅游出游人均花费按文化水平进行分类的数据。试据此用本章介绍的方法将各城市按文化水平进行聚类。

表 18.11　我国 2006 年城镇居民国内旅游出游人均花费情况统计（按文化水平分组）（单位：元/人）

城市	大专及以上	中专及高中	初中	小学	小学以下
北京	1158.8	641.5	567.2	459.3	129.6
天津	914.0	706.9	634.6	558.9	531.1
石家庄	950.3	604.0	451.6	510.1	327.8
太原	1189.2	1188.3	2867.3	781.0	1469.1
呼和浩特	2506.8	1031.5	1894.2	1864.4	286.0
沈阳	499.4	332.8	307.3	214.1	116.0
大连	370.2	261.3	197.7	222.6	88.8
长春	2382.9	2003.9	1089.3	716.1	2200.0
哈尔滨	1670.1	1103.5	925.1	292.4	351.8
上海	1068.6	1242.4	595.6	492.9	222.9
南京	1829.3	988.9	1066.2	1216.6	465.2
无锡	1556.8	851.4	823.0	1602.5	121.0
苏州	5373.1	622.6	527.7	523.7	440.0
杭州	994.9	806.2	681.3	523.8	530.8
青岛	1860.4	1288.0	859.9	265.1	143.4
郑州	989.1	770.5	623.5	563.9	739.8
武汉	1525.5	1044.7	600.9	534.2	51.7
长沙	1844.1	1128.9	640.0	498.0	340.2
广州	820.5	631.0	480.0	510.5	263.5

（续表）

城市	大专及以上	中专及高中	初中	小学	小学以下
深圳	1980.0	1509.6	1859.3	927.1	315.3
银川	1527.7	1223.3	1099.5	1410.2	0
乌鲁木齐	1016.5	825.2	507.9	439.7	254.6

（4）表 18.12 是 2007 年我国 22 个城市城镇居民国内旅游出游人均花费按旅游目的进行分类的数据。试据此用本章介绍的方法将各城市按旅游目的进行聚类。

表 18.12　我国 2007 年城镇居民国内旅游出游人均花费情况统计（按旅游目的分组）（单位：元/人）

城市	观光游览	探亲访友	商务	公务会议	度假休闲
北京	911.5	889.7	1025.2	1909.6	671.5
天津	819.1	618.5	833.2	1324.0	822.4
石家庄	653.9	349.8	905.0	1930.5	789.9
太原	1689.6	871.1	0	2898.5	1007.9
呼和浩特	2876.1	1067.6	453.2	2810.5	1162.3
沈阳	411.3	315.0	1842.0	609.7	303.2
大连	307.7	316.6	218.8	1286.1	138.2
长春	1826.2	844.7	3313.0	4222.3	1684.1
哈尔滨	1066.1	1008.2	2955.5	2687.0	1249.6
上海	737.8	2548.2	4786.1	780.5	783.4
南京	1236.6	1180.6	1876.3	976.3	2420.9
无锡	1028.9	388.1	2204.6	0	2727.1
苏州	1940.2	464.0	11843.3	762.3	550.0
杭州	987.9	518.3	343.8	1860.1	387.3
青岛	1514.3	829.5	4390.8	2950.4	1202.9
郑州	823.8	681.7	335.5	0	863.3
武汉	1341.6	964.2	4860.2	2427.9	879.4
长沙	1585.3	1379.9	4369.4	2326.5	553.3
广州	721.8	719.7	233.2	662.8	358.5
深圳	2466.9	1086.1	1013.5	1466.9	1885.3
银川	1577.7	965.8	2824.8	915.2	1333.6
乌鲁木齐	553.2	875.0	2418.9	3584.4	364.2

（5）根据表 18.13 中 26 个著名的风景区的相关资料，按照相关特征变量对景区名称进行聚类分析。

表 18.13　国家级风景名胜区数据统计

风景名胜区名称	风景名胜区面积/平方千米	游人量/万人次	景区资金收入/万元	景区资金支出/万元
西岭雪山	483	58	9150	4532
青城山	150	350	16321	3845
龙门山	81	49	940	715
天台山	106	20	2255	1508
剑门蜀道	597	196	1844	1359
白龙湖	482	1	210	190
峨眉山	138	668	364 294	78 534
蜀南竹海	120	58	29 497	21 053
石海洞乡	156	24	1651	1660

（续表）

风景名胜区名称	风景名胜区面积/平方千米	游人量/万人次	景区资金收入/万元	景区资金支出/万元
云雾山	775	30	3300	1500
黄龙寺	2060	440	72 951	38 649
四姑娘山	480	11	798	4558
泸沽湖	161	11	6400	1600
螺髻山	2240	8	597	270
红枫湖	200	14	820	780
赤水	630	70	3614	5311
龙宫	60	105	4505	3683
黄果树	163	403	142 374	20 800
紫云格	57	23	402	316
九龙洞	56	7	118	61
马岭河峡	450	10	428	428
织金洞	307	28	1204	988
水舞阳河	625	134	1862	1862
荔波樟江	275	70	4230	1250
都匀斗蓬	267	4	480	386
昆明滇池	685	456	9860	8200

第 19 章　Stata 在经济增长分析中的应用

近年来，党和政府高度重视经济增长方式的有效转变问题。中国共产党的十八大报告指出，在当代中国，以科学发展为主题，以加快转变经济发展方式为主线，是关系我国发展全局的战略抉择。

关于经济增长方式的分类，目前比较流行也比较常用的做法是把它分为粗放型增长和集约型增长两类。其中，粗放型增长是在效率没有明显提高的情况下，主要依靠量的积累，依靠更多包括资本、劳动力等资源的投入来实现经济增长和经济总量增加的增长方式，这也是经济体在发展初始通常需要经历的一个阶段。与粗放型增长不同的是，集约型增长非常注重技术的改进与升级，注重资源利用效率的提升，注重生产效率的有效提高，强调质的方面，强调在不依靠更多包括资本、劳动力等资源的投入前提下，通过提高投入产出比来实现经济增长和经济总量增加。通常所说的经济增长方式的转变就是经济增长方式由粗放型增长方式向集约型增长方式的转变。本章就以实例的形式来介绍一下 Stata 14.0 在经济增长分析中的应用。

2012 年，济南市面临着经济形势复杂严峻、社会矛盾日益凸显、改革发展稳定的压力不断加大的重重困难。在这种情况下，市委、市政府提出了要坚持以科学发展观为指导，牢牢把握科学发展主题的指导思想，并把发展实体经济、建设美丽泉城、优化发展环境、创新社会管理 4 方面工作作为经济社会发展的重中之重，作为各级党委、政府全力突破的主攻方向。根据发展是硬道理、发展是解决所有问题的关键的指导思想，做好这 4 方面工作的基础和根本就在于把加快转变经济增长方式作为主线。加快经济增长方式转变对于推动济南市又好又快可持续发展意义重大。本章的研究目的在于通过实例分析来探索济南市目前经济增长方式的具体情况。

19.1　数据来源与研究思路

本章所用的数据包括济南市 1994－2010 年地区生产总值、固定资产投资、年底就业人数、财政科技投入等时间序列数据。所有数据均取自历年《济南统计年鉴》。数据的 Excel 形式如表 19.1 所示。

表 19.1　案例数据

年份	地区生产总值/亿元	固定资产投资/亿元	年底就业人数/万人	财政科技投入/万元
1994	372	73.2	304	1432
1995	473.52	112.76	324.2	2307
1996	569.252	150.8	332.3	3634
1997	664.984	181.24	337.4	6123
1998	760.716	220.69	341.6	7687
1999	856.448	270.42	344.5	12650

（续表）

年份	地区生产总值/亿元	固定资产投资/亿元	年底就业人数/万人	财政科技投入/万元
2000	952.18	305.95	347.4	13 027
2001	1 066.16	344.15	350.1	18 659
2002	1 200.83	404.69	352.7	20 184
2003	1 365.33	504.89	355.3	14 590
2004	1 618.87	651.3	358.5	20 251
2005	1 846.28	857	360	22 383
2006	2 161.53	1 016.77	361.8	27 537
2007	2 500.14	1 151.7	364.3	40 516
2008	3 006.77	1 415.33	367.4	45 062
2009	3 340.91	1 655.37	372.3	52 625
2010	3 910.53	1 987.44	373.7	62 138

本数据为时间序列数据，研究思路是：首先对数据进行描述性分析，并绘制变量的时间序列趋势图，简明扼要地分析一下数据特征，并进行了相关性检验，探索变量之间的相关关系，然后对数据中各个时间序列变量采用多种方法进行单位根检验，综合分析其平稳性，使用回归分析方法探索平稳变量之间的关系，并使用迹检验这种协整检验的方式对非平稳数据进行协整检验，综合分析其长期均衡关系，又对两个非平稳变量进行了格兰杰因果关系检验，探讨变量之间的格兰杰因果关系，最后建立了相应的误差修正模型，并提出了研究结论。

19.2　描述性分析

	下载资源:\video\chap19\…
	下载资源:\sample\chap19\案例19.dta

本案例的数据变量都是定距变量，通过进行定距变量的基本描述性统计，可以得到数据的概要统计指标，包括平均值、最大值、最小值、标准差、百分位数、中位数、偏度系数和峰度系数等。通过获得这些指标，可以从整体上对拟分析的数据进行宏观的把握，为后续进行更深入的数据分析做好必要准备。

19.2.1　Stata 分析过程

在用 Stata 进行分析之前，要把数据录入到 Stata 中。本例中有 5 个变量，分别为年份、地区生产总值、固定资产投资、年底就业人数和财政科技投入。我们把年份变量设定为 year，把地区生产总值变量设定为 gdp，把固定资产投资变量设定为 invest，把年底就业人数变量设定为 labor，把财政科技投入变量设定为 scientific，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作在第 1 章中已有详细讲述。录入完成后数据如图 19.1 所示。

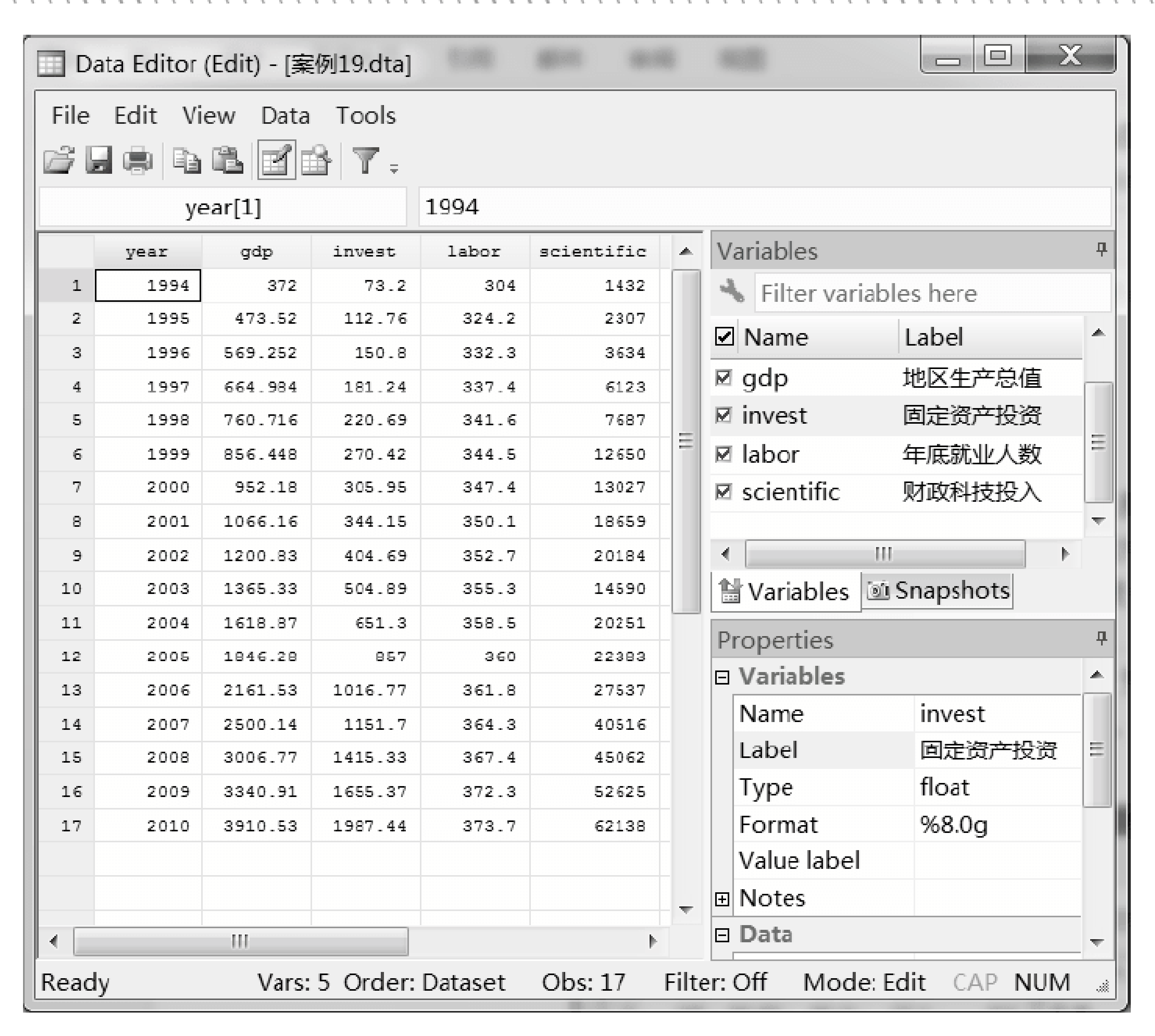


图 19.1　案例 19 数据

先做一下数据保存，然后开始展开分析，值得说明的是，本例中需要对各个时间序列变量数据进行对数标准化处理，一方面我们可以消除数据异方差的影响，使数据更适合深入分析，并且使数据更具实际意义，另一方面可以研究变量之间的弹性关系。在没有进行对数变换之前，变量之间的联动关系表现为自变量的变动引起因变量变动的程度，在进行对数变换之后，变量的联动关系就表现为自变量变动的百分比引起因变量变动的百分比的程度。此外这种处理模式也契合了经济增长的理论的经典模型之一：柯布-道格拉斯生产函数模型。该模型常用的表述形式是：

$$\ln Yt = \alpha \ln Kt + \beta \ln Lt + \gamma \ln Tt + \ln At + \mu$$

其中，Yt、Kt、Lt、Tt 分别表示地区生产总值、固定资产投资、年底就业人数和财政科技投入。α、β 和 γ 分别表示固定资产投资、年底就业人数和财政科技投入的产出弹性，lnAt 为常数项，而μ是随机误差项。

描述性分析的步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入如下命令。

- generate lgdp=ln(gdp)：本命令旨在对变量“gdp”进行对数变换。
- generate linvest=ln(invest)：本命令旨在对变量“invest”进行对数变换。
- generate llabor=ln(labor)：本命令旨在对变量“labor”进行对数变换。
- generate lscientific=ln(scientific)：本命令旨在对变量“scientific”进行对数变换。
- summarize　gdp invest labor scientific lgdp linvest llabor lscientific,detail：本命令旨在对地区生产总值、固定资产投资、年底就业人数和财政科技投入等变量以及它们的对数标准化变量进行描述性分析。

03 设置完毕后，按键盘上的回车键，等待输出结果。

19.2.2　结果分析

在 Stata 14.0 主界面的结果窗口可以看到如图 19.2~图 19.9 所示的分析结果。

1．数据标准化处理结果

选择“Data” | “Data Editor” | “Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 19.2 所示的 lgdp 数据。lgdp 数据是对数据 gdp 进行对数变换处理的结果。

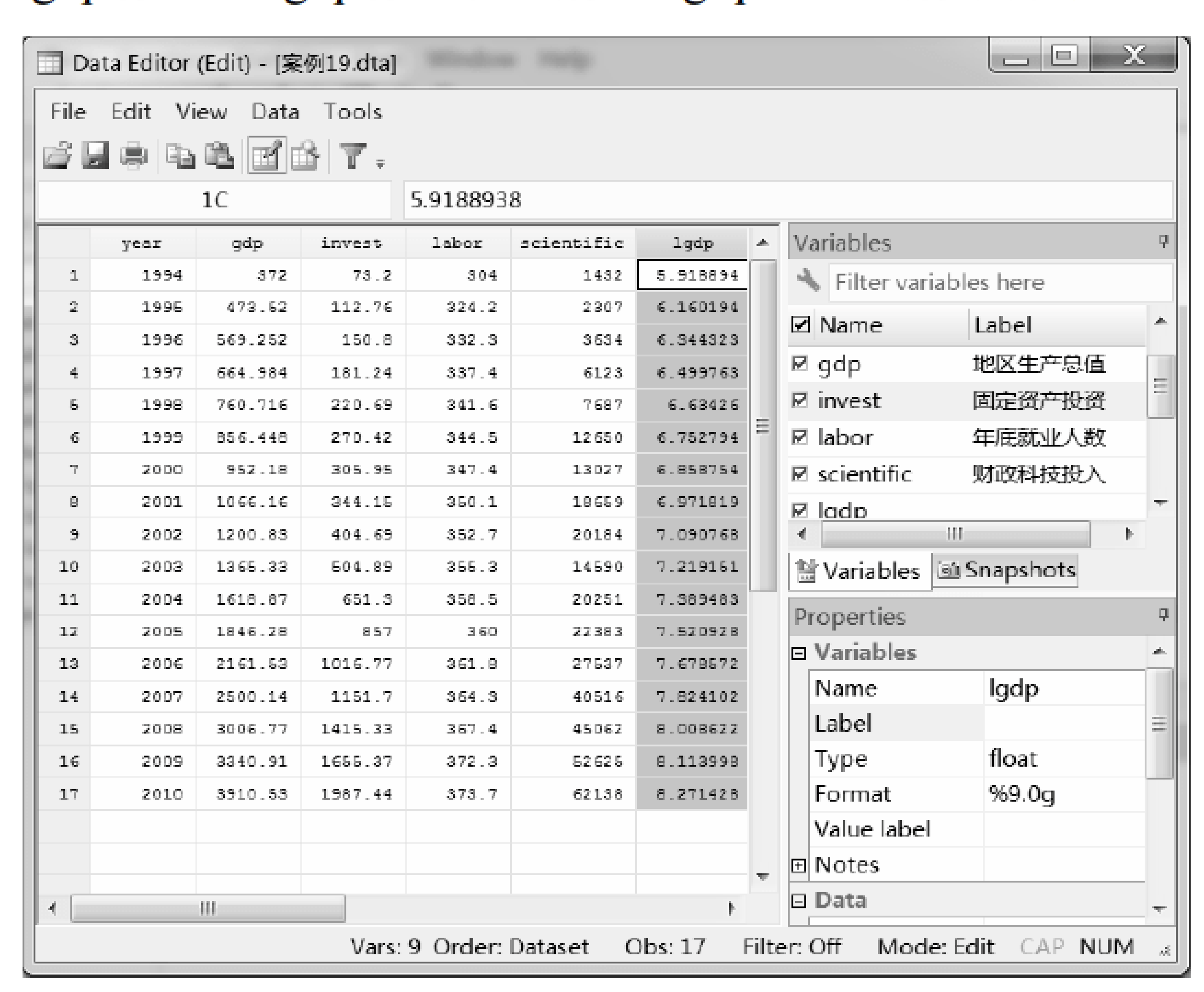


图 19.2　数据标准化处理分析结果 1

选择“Data” | “Data Editor” | “Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 19.3 所示的 linvest 数据。linvest 数据是对数据 invest 进行对数变换处理的结果。

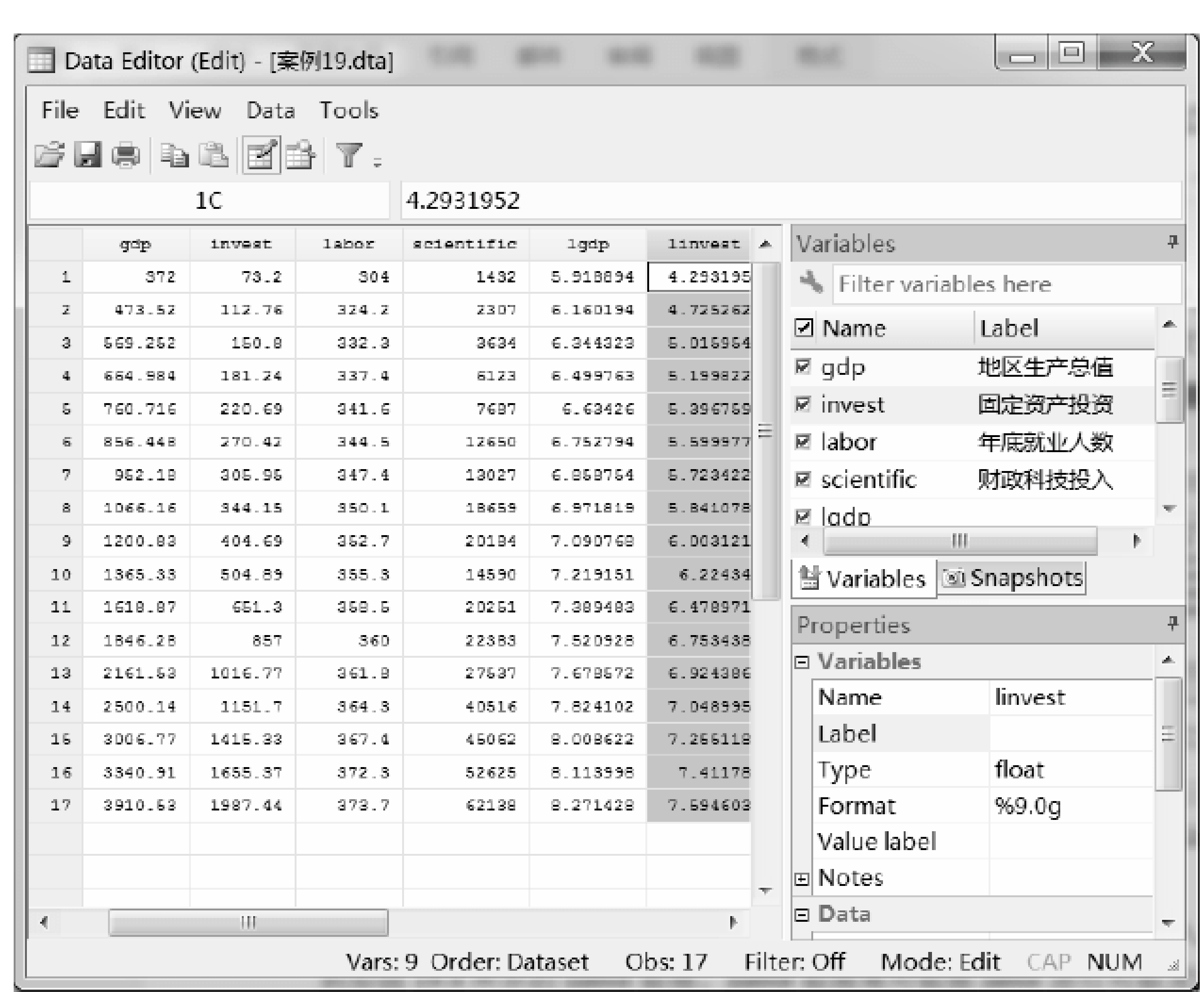


图 19.3　数据标准化处理分析结果 2

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 19.4 所示的 llabor 数据。llabor 数据是对数据 labor 进行对数变换处理的结果。

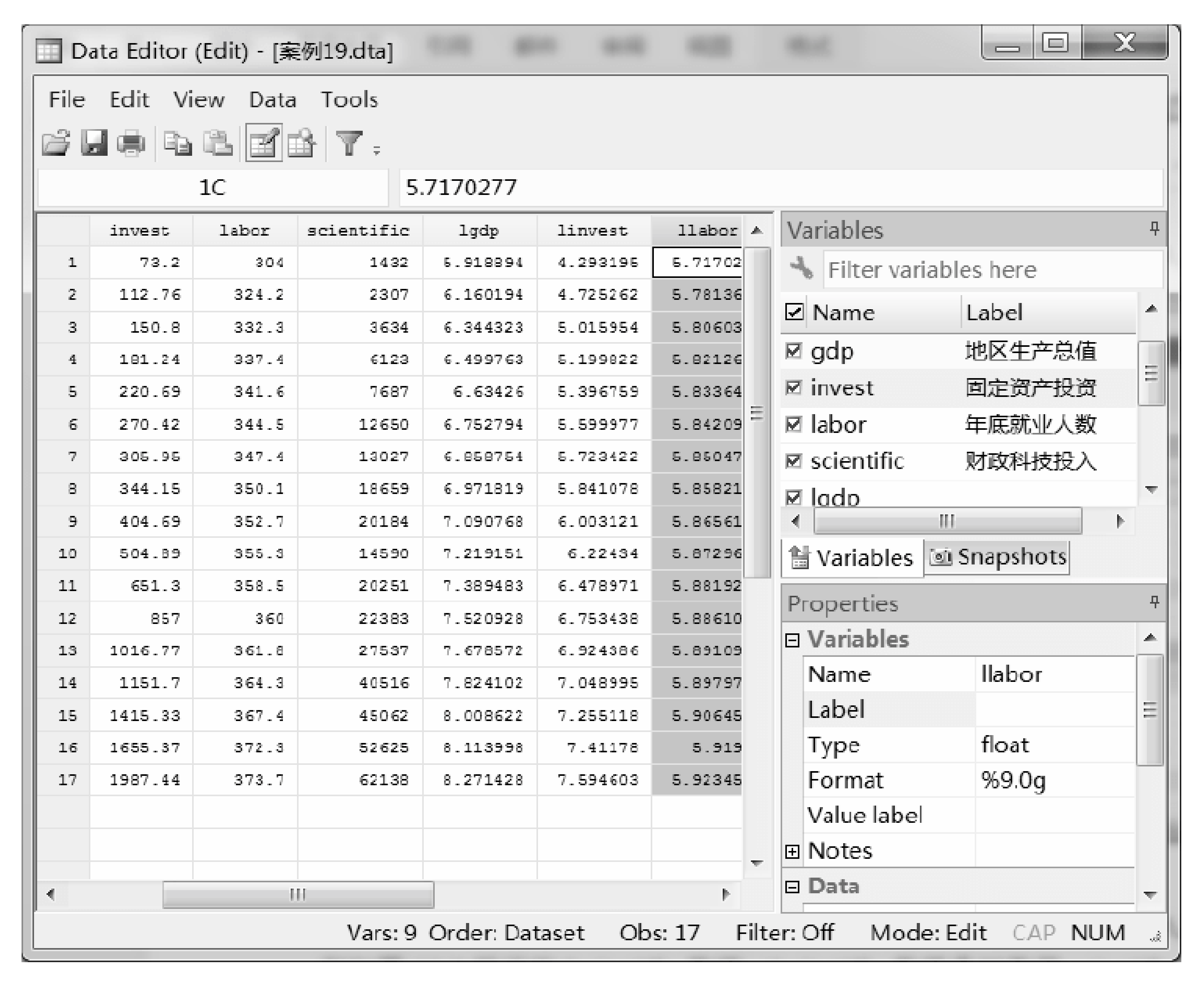


图 19.4　数据标准化处理分析结果 3

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 19.5 所示的 lscientific 数据。lscientific 数据是对数据 scientific 进行对数变换处理的结果。

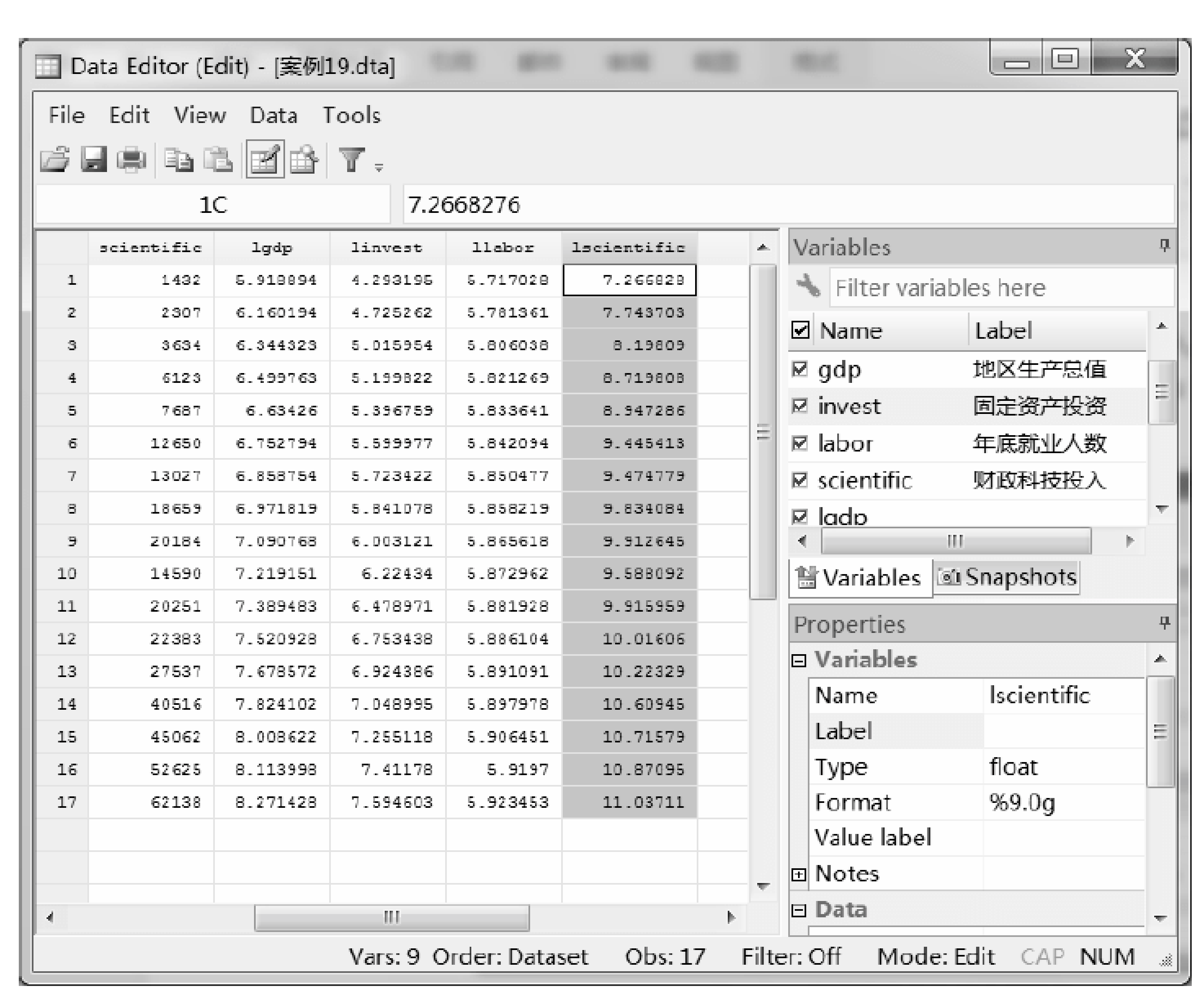


图 19.5　数据标准化处理分析结果 4

2. 描述性分析结果

图 19.6~图 19.9 给出了地区生产总值、固定资产投资、年底就业人数和财政科技投入等变量以及其对数标准化变量的描述性分析结果。

地区生产总值

	Percentiles	Smallest		
1%	372	372		
5%	372	473.52		
10%	473.52	569.252	Obs	17
25%	760.716	664.984	Sum of Wgt.	17
50%	1200.83		Mean	1568.615
		Largest	Std. Dev.	1072.617
75%	2161.53	2500.14		
90%	3340.91	3006.77	Variance	1150507
95%	3910.53	3340.91	Skewness	.8537391
99%	3910.53	3910.53	Kurtosis	2.567163

固定资产投资

	Percentiles	Smallest		
1%	73.2	73.2		
5%	73.2	112.76		
10%	112.76	150.8	Obs	17
25%	220.69	181.24	Sum of Wgt.	17
50%	404.69		Mean	664.9235
		Largest	Std. Dev.	587.491
75%	1016.77	1151.7		
90%	1655.37	1415.33	Variance	345145.6
95%	1987.44	1655.37	Skewness	.9527296
99%	1987.44	1987.44	Kurtosis	2.70205

图 19.6　描述性分析结果图 1

年底就业人数

	Percentiles	Smallest		
1%	304	304		
5%	304	324.2		
10%	324.2	332.3	Obs	17
25%	341.6	337.4	Sum of Wgt.	17
50%	352.7		Mean	349.8529
		Largest	Std. Dev.	18.15558
75%	361.8	364.3		
90%	372.3	367.4	Variance	329.6251
95%	373.7	372.3	Skewness	-.9094851
99%	373.7	373.7	Kurtosis	3.502769

财政科技投入

	Percentiles	Smallest		
1%	1432	1432		
5%	1432	2307		
10%	2307	3634	Obs	17
25%	7687	6123	Sum of Wgt.	17
50%	18659		Mean	21812.06
		Largest	Std. Dev.	18199.12
75%	27537	40516		
90%	52625	45062	Variance	3.31e+08
95%	62138	52625	Skewness	.8945362
99%	62138	62138	Kurtosis	2.726649

图 19.7　描述性分析结果图 2

lgdp

	Percentiles	Smallest		
1%	5.918894	5.918894		
5%	5.918894	6.160194		
10%	6.160194	6.344323	Obs	17
25%	6.63426	6.499763	Sum of Wgt.	17
50%	7.090768		Mean	7.132815
		Largest	Std. Dev.	.7055021
75%	7.678572	7.824102		
90%	8.113998	8.008622	Variance	.4977332
95%	8.271428	8.113998	Skewness	.0009106
99%	8.271428	8.271428	Kurtosis	1.94138

linvest

	Percentiles	Smallest		
1%	4.293195	4.293195		
5%	4.293195	4.725262		
10%	4.725262	5.015954	Obs	17
25%	5.396759	5.199822	Sum of Wgt.	17
50%	6.003121		Mean	6.08766
		Largest	Std. Dev.	.9860055
75%	6.924386	7.048995		
90%	7.41178	7.255118	Variance	.9722068
95%	7.594603	7.41178	Skewness	-.1062487
99%	7.594603	7.594603	Kurtosis	1.947088

图 19.8　描述性分析结果图 3

llabor

	Percentiles	Smallest		
1%	5.717028	5.717028		
5%	5.717028	5.781361		
10%	5.781361	5.806038	Obs	17
25%	5.833641	5.821269	Sum of Wgt.	17
50%	5.865618		Mean	5.856201
		Largest	Std. Dev.	.05328
75%	5.891091	5.897978		
90%	5.9197	5.906451	Variance	.0028388
95%	5.923453	5.9197	Skewness	-1.042006
99%	5.923453	5.923453	Kurtosis	3.820275

lscientific

	Percentiles	Smallest		
1%	7.266828	7.266828		
5%	7.266828	7.743703		
10%	7.743703	8.19809	Obs	17
25%	8.947286	8.719808	Sum of Wgt.	17
50%	9.834084		Mean	9.559961
		Largest	Std. Dev.	1.084743
75%	10.22329	10.60945		
90%	10.87095	10.71579	Variance	1.176667
95%	11.03711	10.87095	Skewness	-.6424312
99%	11.03711	11.03711	Kurtosis	2.570768

图 19.9　描述性分析结果图 4

在如图 19.6~图 19.9 所示的分析结果中，可以得到很多信息。此处限于篇幅不再针对各个变量一一展开说明，以变量 lscientific 为例进行解释。

- 百分位数(Percentiles)：可以看出变量 lscientific 的第 1 个四分位数(25%)是 8.947286，第 2 个四分位数（50%）是 9.834084。

- 4 个最小值（Smallest）：变量 lscientific 最小的 4 个数据值分别是 7.266828、7.743703、8.19809、8.719808。
- 4 个最大值（Largest）：变量 lscientific 最大的 4 个数据值分别是 10.60945、10.71579、10.87095、11.03711。
- 平均值（Mean）和标准差（Std. Dev）：变量 lscientific 的平均值为 9.559961，标准差是 1.084743。
- 偏度（Skewness）和峰度（Kurtosis）：变量 lscientific 的偏度为-0.6424312，为负偏度但不大。变量 lscientific 的峰度为 2.570768，有一个比正态分布略短的尾巴。

从上面的描述性分析结果中可以看出，所有数据中没有极端数据，数据间的量纲差距也在可接受范围之内，可以进入下一步的分析过程。

19.3 时间序列趋势图

我们通过绘制时间序列趋势图操作可以迅速看出数据的变化特征，为后续更加精确地判断或者选择合适的模型做好必要准备。

19.3.1 Stata 分析过程

时间序列趋势图分析的步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入如下命令。

- tsset year：本命令旨在把数据定义为时间序列，时间变量为“year”。
- twoway(line gdp year)：本命令旨在绘制变量“gdp”随时间变量“year”变动的时间趋势图。
- twoway(line invest year)：本命令旨在绘制变量“invest”随时间变量“year”变动的时间趋势图。
- twoway(line labor year)：本命令旨在绘制变量“labor”随时间变量“year”变动的时间趋势图。
- twoway(line scientific year)：本命令旨在绘制变量“scientific”随时间变量“year”变动的时间趋势图。
- twoway(line lgdp year)：本命令旨在绘制变量“lgdp”随时间变量“year”变动的时间趋势图。
- twoway(line linvest year)：本命令旨在绘制变量“linvest”随时间变量“year”变动的时间趋势图。
- twoway(line llabor year)：本命令旨在绘制变量“llabor”随时间变量“year”变动的时间趋势图。
- twoway(line lscientific year)：本命令旨在绘制变量“lscientific”随时间变量“year”变

动的时间趋势图。

- twoway(line d.lgdp year)：本命令旨在绘制变量“d.lgdp”随时间变量“year”变动的时间趋势图。
- twoway(line d.linvest year)：本命令旨在绘制变量“d.linvest”随时间变量“year”变动的时间趋势图。
- twoway(line d.llabor year)：本命令旨在绘制变量“d.llabor”随时间变量“year”变动的时间趋势图。
- twoway(line d.lscientific year)：本命令旨在绘制变量“d.lscientific”随时间变量“year”变动的时间趋势图。

03 设置完毕后，按键盘上的回车键，等待输出结果。

19.3.2　结果分析

在 Stata 14.0 主界面的结果窗口可以看到如图 19.10~图 19.22 所示的分析结果。

图 19.10 显示的是我们把年份作为日期变量对数据进行时间定义的结果。

```
. tsset year
        time variable:  year, 1994 to 2010
                delta:  1 unit
```

图 19.10　时间序列趋势图分析结果图 1

从上述分析结果中，可以看到时间变量是年份（year），区间范围是从 1994~2010，间距为 1。

图 19.11 显示的是变量地区生产总值随时间的变动趋势。

从上述分析结果中，可以看到变量地区生产总值具有明显、稳定的长期增长趋势。

图 19.12 显示的是变量固定资产投资随时间的变动趋势。

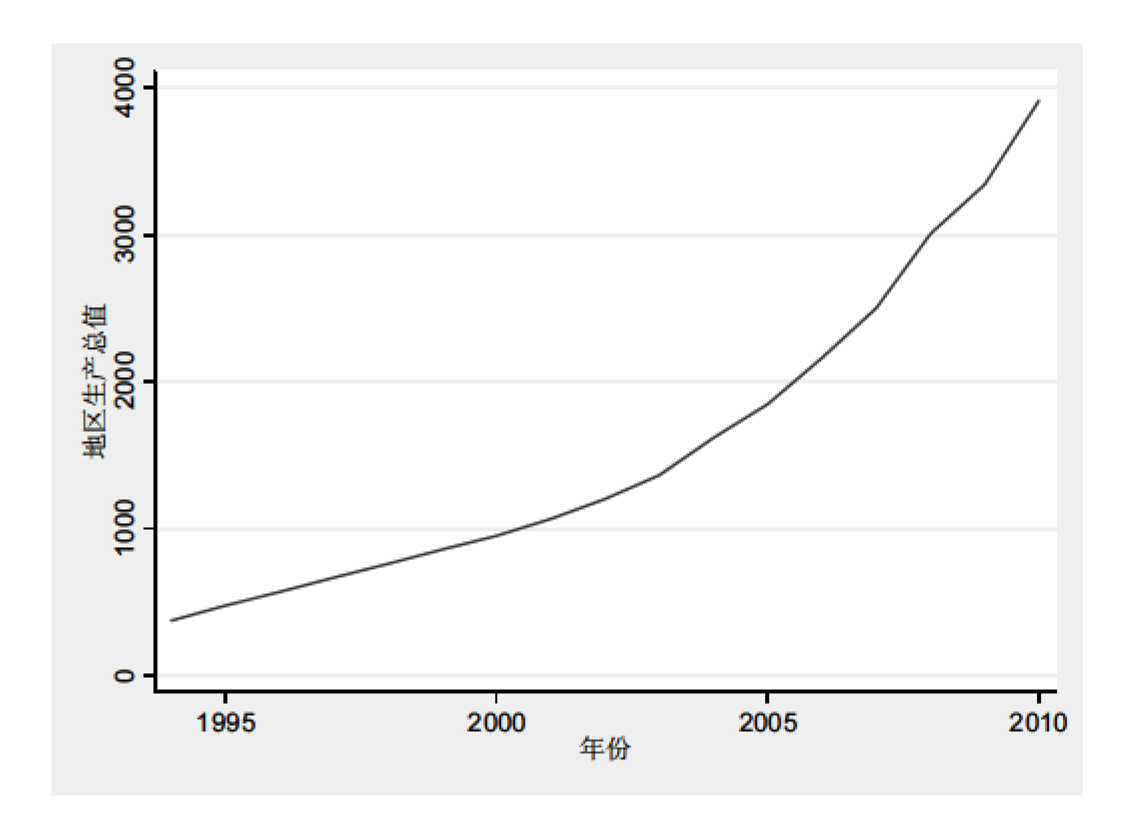


图 19.11　时间序列趋势图分析结果图 2

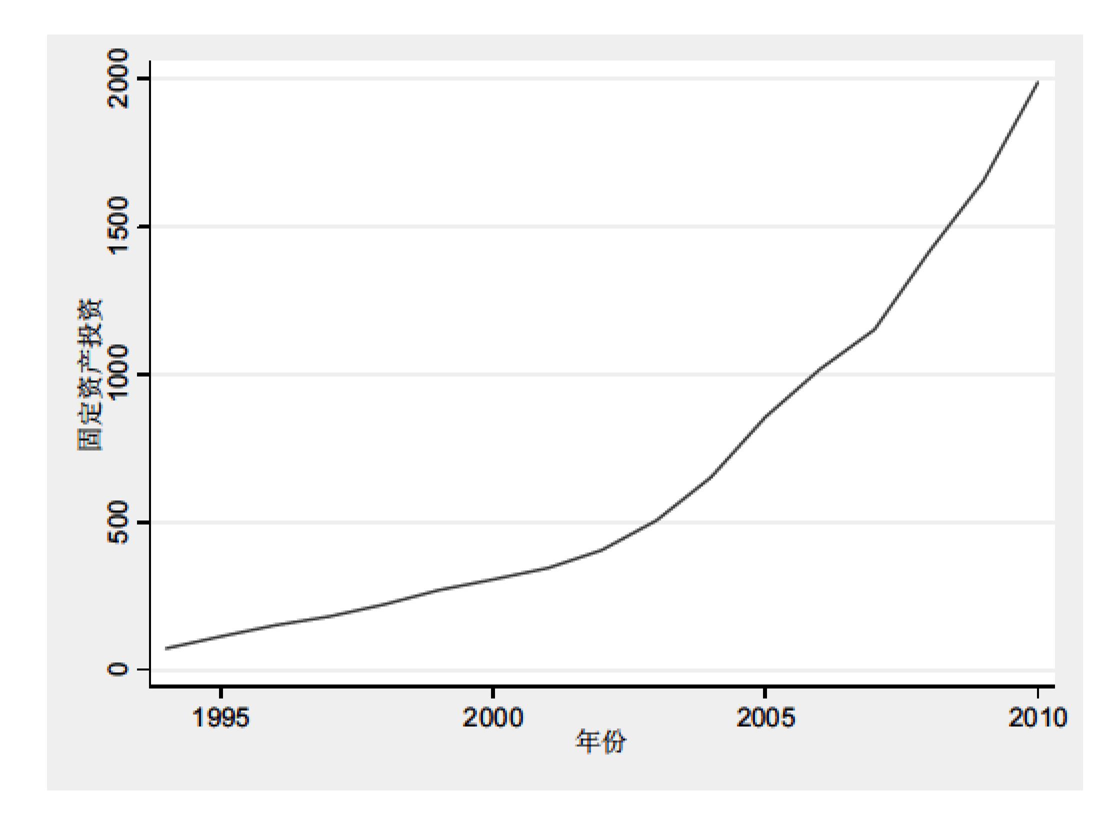


图 19.12　时间序列趋势图分析结果图 3

从上述分析结果中，可以看到变量固定资产投资具有明显、稳定的长期增长趋势。

图 19.13 显示的是变量年底就业人数随时间的变动趋势。从分析结果中，可以看到变量年底就业人数具有明显、稳定的向上增长趋势。

图 19.14 显示的是变量财政科技投入随时间的变动趋势。

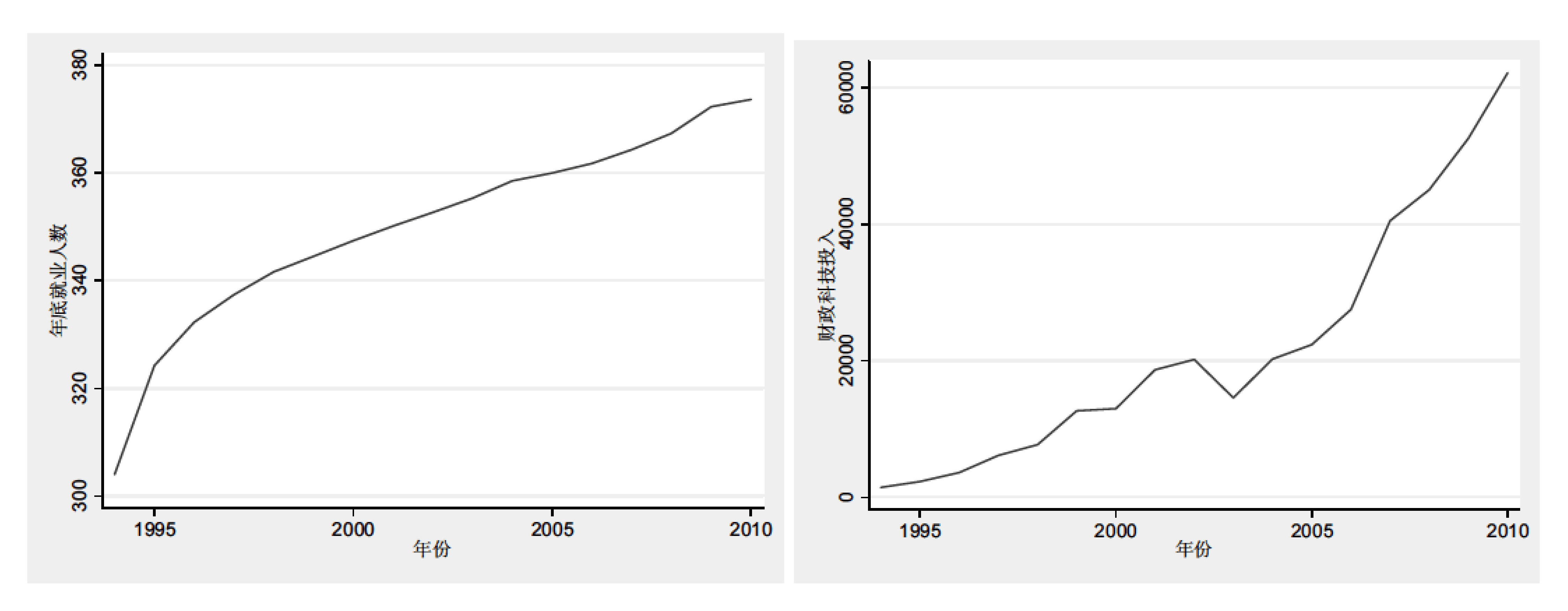


图 19.13　时间序列趋势图分析结果图 4

图 19.14　时间序列趋势图分析结果图 5

从上述分析结果中，可以看到变量财政科技投入具有明显、稳定的长期变动趋势。

图 19.15 显示的是变量地区生产总值的对数值随时间的变动趋势。从分析结果中，可以看到变量地区生产总值的对数值具有明显、稳定的长期增长趋势。

图 19.16 显示的是变量固定资产投资的对数值随时间的变动趋势。

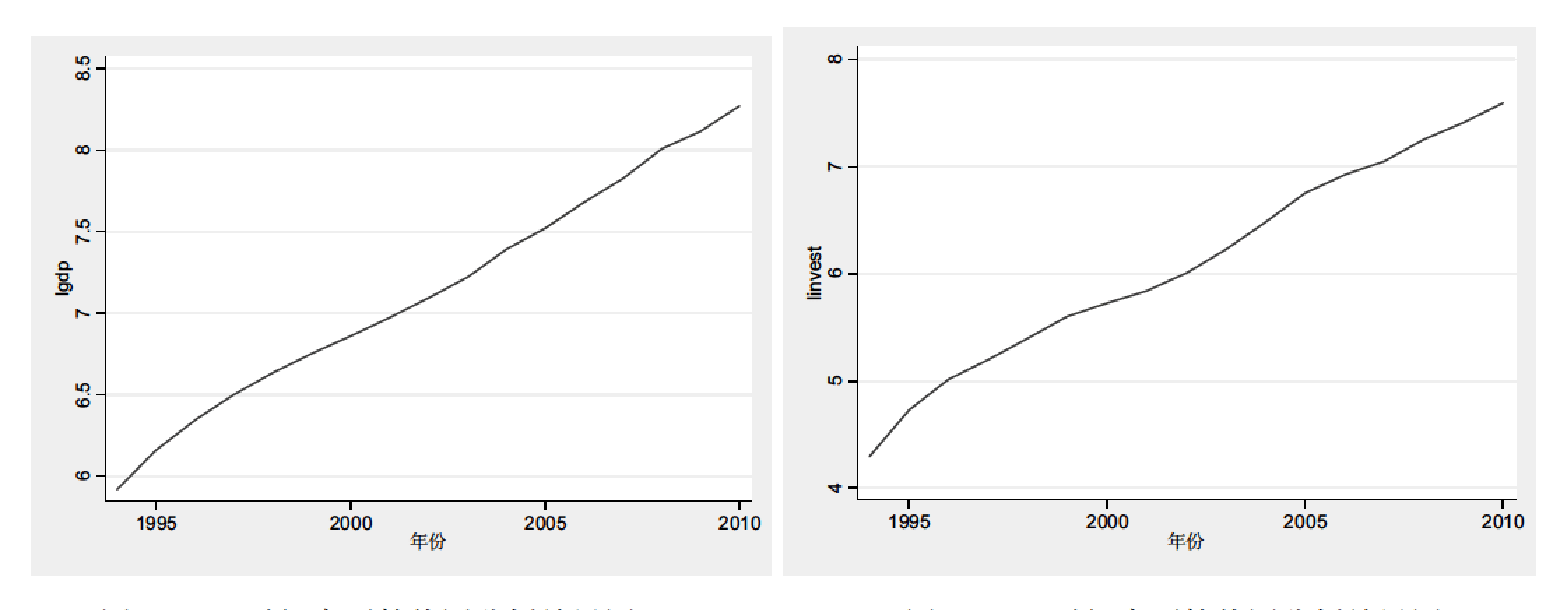


图 19.15　时间序列趋势图分析结果图 6

图 19.16　时间序列趋势图分析结果图 7

从上述分析结果中，可以看到变量固定资产投资的对数值具有明显、稳定的长期增长趋势。

图 19.17 显示的是变量年底就业人数的对数值随时间的变动趋势。从分析结果中，可以看到变量年底就业人数的对数值具有明显、稳定的向上增长趋势。

图 19.18 显示的是变量财政科技投入的对数值随时间的变动趋势。

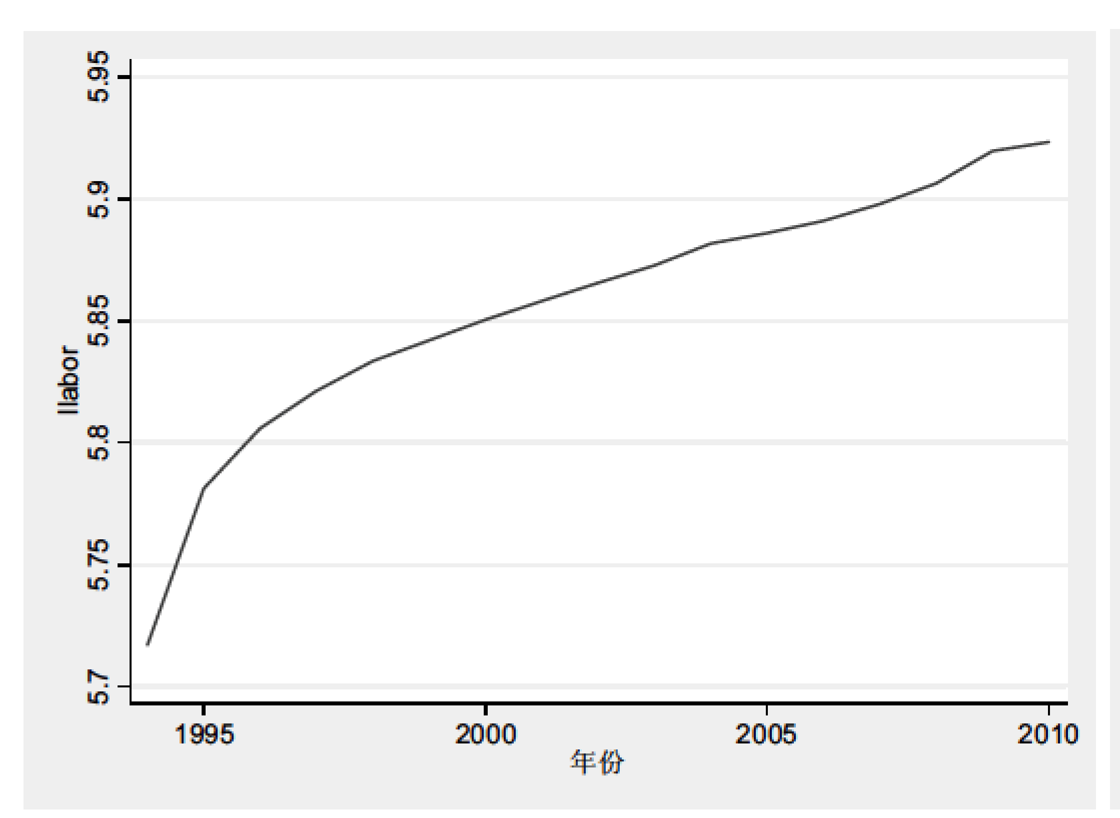


图 19.17　时间序列趋势图分析结果图 8

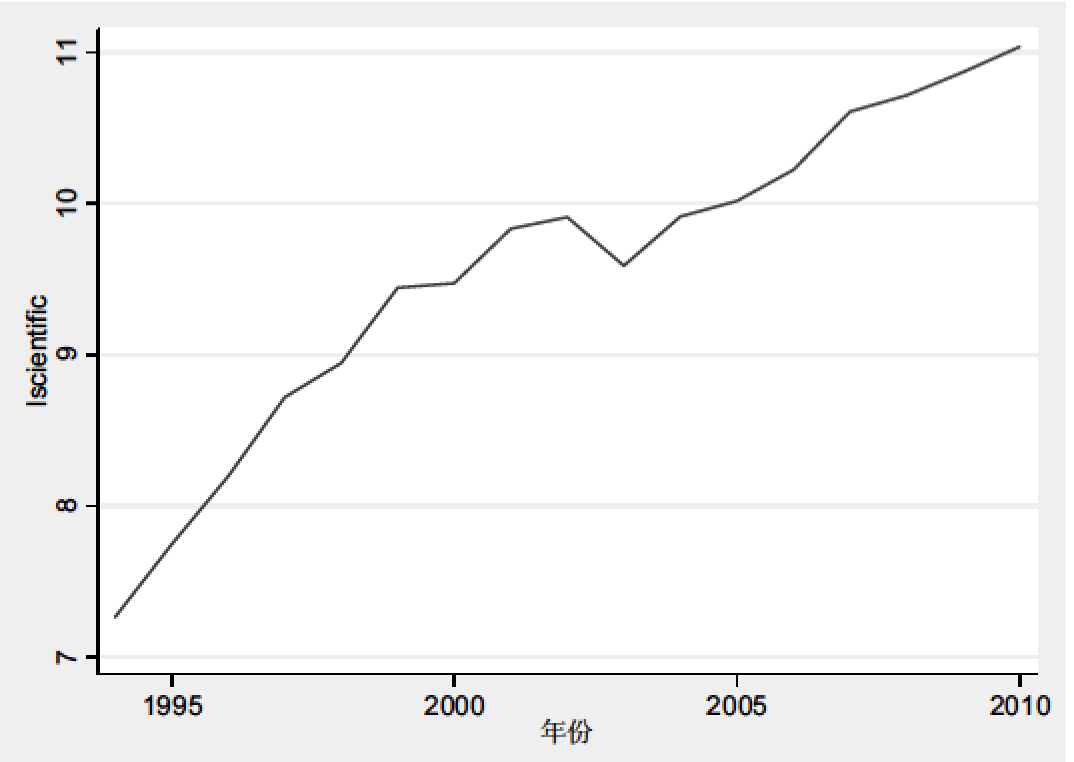


图 19.18　时间序列趋势图分析结果图 9

从上述分析结果中，可以看到变量财政科技投入的对数值具有明显、稳定的长期变动趋势。

图 19.19 显示的是变量地区生产总值的对数值的一阶差分值随时间的变动趋势，从分析结果中，可以看到变量地区生产总值的对数值的一阶差分值没有明显、稳定的长期增长趋势。

图 19.20 显示的是变量固定资产投资的对数值的一阶差分值随时间的变动趋势。

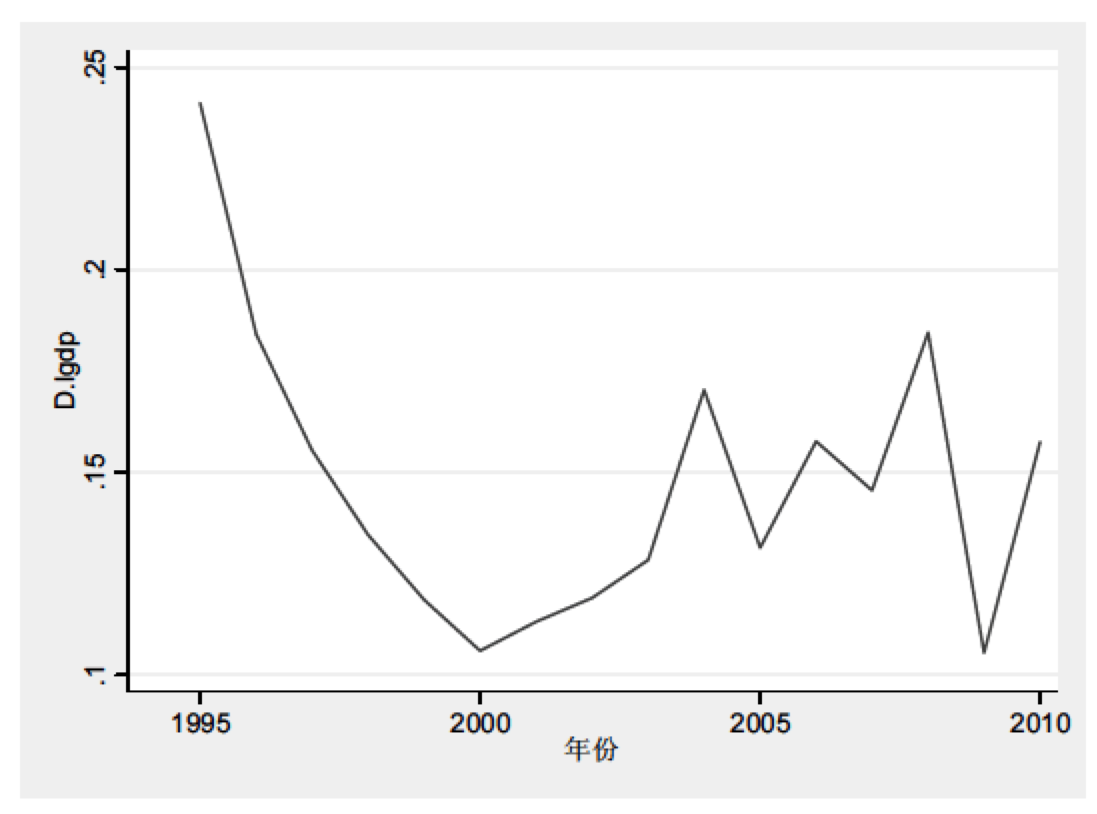


图 19.19　时间序列趋势图分析结果图 10

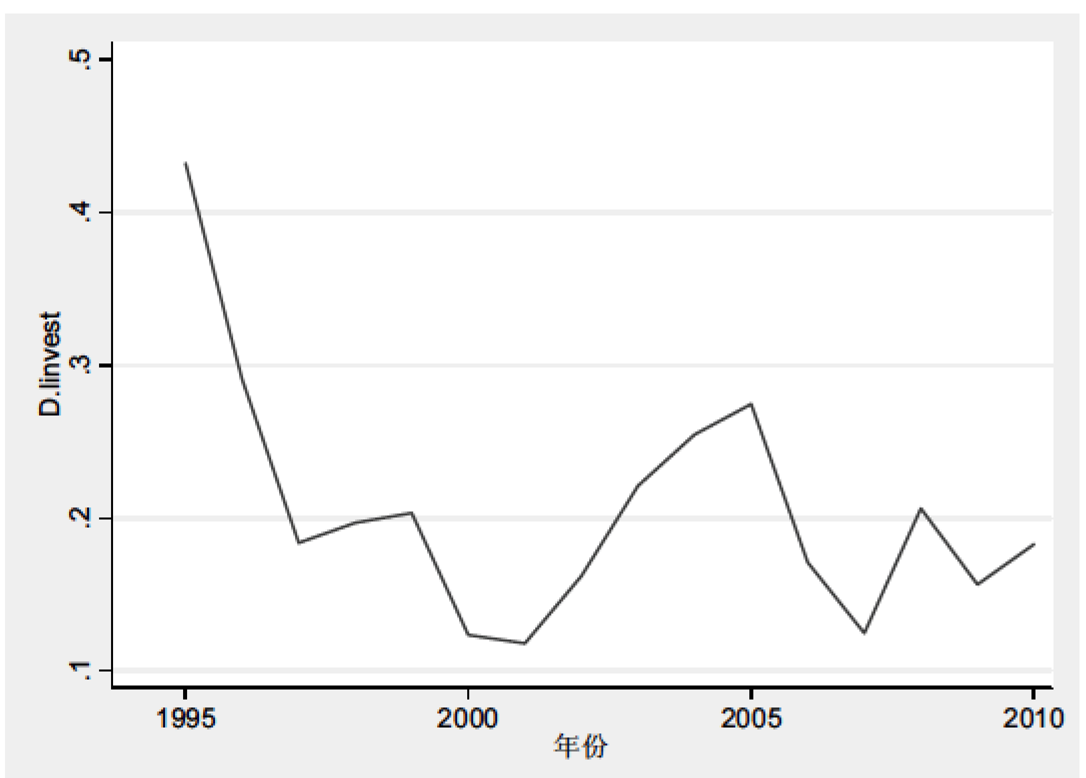


图 19.20　时间序列趋势图分析结果图 11

从上述分析结果中，可以看到变量固定资产投资的对数值的一阶差分值没有明显、稳定的长期增长趋势。

图 19.21 显示的是变量年底就业人数的对数值的一阶差分值随时间的变动趋势，从分析结果中，可以看到变量年底就业人数的对数值的一阶差分值没有明显、稳定的向上增长趋势。

图 19.22 显示的是变量财政科技投入的对数值的一阶差分值随时间的变动趋势。

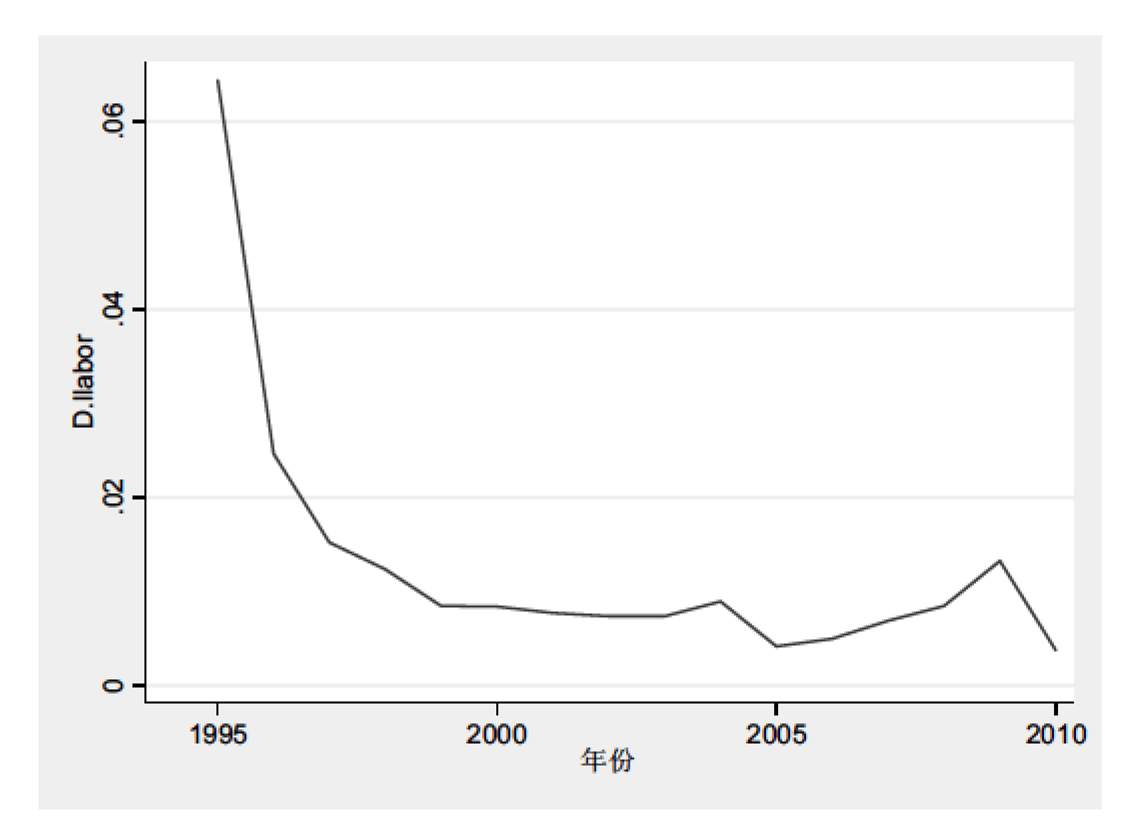


图 19.21　时间序列趋势图分析结果图 12

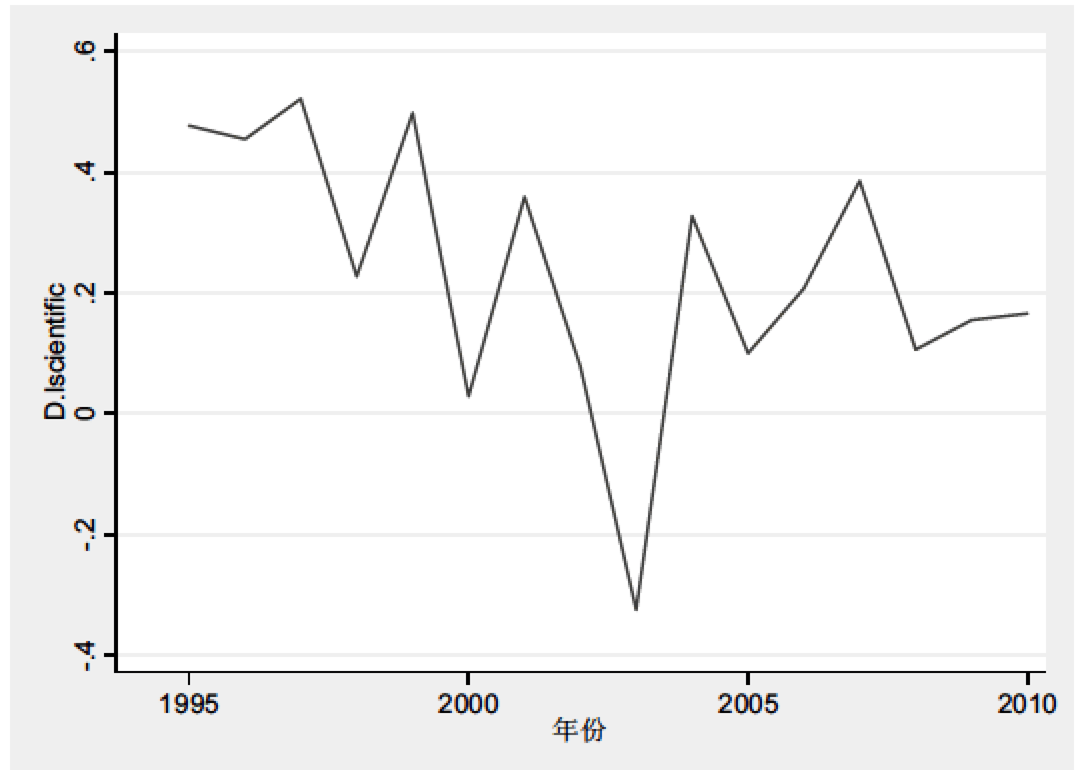


图 19.22　时间序列趋势图分析结果图 13

从上述分析结果中，可以看到变量财政科技投入的对数值的一阶差分值没有明显、稳定的长期变动趋势。

综上所述，通过绘制时间序列趋势图发现变量地区生产总值、固定资产投资、年底就业人数、财政科技投入的值以及其对数标准化的值都是有明显、稳定的向上增长趋势的，而变量地区生产总值、固定资产投资、年底就业人数、财政科技投入的对数值的一阶差分值是没有明显、稳定的时间趋势的。这些结论将会在后续的操作命令中被用到。

19.4　相关性分析

相关分析是不考虑变量之间的因果关系而只研究分析变量之间的相关关系的一种统计分析方法，通过该步操作可以判断出变量之间的相关性，从而考虑是否有必要进行后续分析或者增加新的变量等。

19.4.1　Stata 分析过程

相关性分析的步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入如下命令：

- correlate　gdp invest labor scientific,covariance
- correlate　lgdp linvest llabor lscientific,covariance
- correlate　gdp invest labor scientific
- correlate　lgdp linvest llabor lscientific
- pwcorr gdp invest labor scientific,sidak sig star(0.01)
- pwcorr lgdp linvest llabor lscientific,sidak sig star(0.01)

03 设置完毕后，按键盘上的回车键，等待输出结果。

19.4.2 结果分析

在 Stata 14.0 主界面的结果窗口可以看到如图 19.23~图 19.28 所示的分析结果。

图 19.23 展示的是变量地区生产总值、固定资产投资、年底就业人数和财政科技投入之间的方差-协方差矩阵。

```
. correlate  gdp invest labor scientific,covariance
(obs=17)

             |      gdp   invest    labor scient~c
-------------+------------------------------------
         gdp |  1.2e+06
      invest |   628790   345146
       labor |    16633  8832.29  329.625
  scientific |  1.9e+07  1.0e+07   280244  3.3e+08
```

图 19.23 相关性分析结果图 1

从上述分析结果中，可以看到地区生产总值的方差是 1.2e+06，固定资产投资的方差是 345146，年底就业人数的方差是 329.625，财政科技投入的方差是 3.3e+08，地区生产总值与固定资产投资之间的协方差是 628790，地区生产总值与年底就业人数之间的协方差是 16633，地区生产总值与财政科技投入之间的协方差是 1.9e+07，固定资产投资与年底就业人数之间的协方差是 8832.29，固定资产投资与财政科技投入之间的协方差是 1.0e+07，财政科技投入与年底就业人数之间的协方差是 280244。可以发现变量之间的方差差别是非常大的，我们对数据进行对数变换处理是非常有必要，也是非常有意义的。

图 19.24 展示的是变量地区生产总值、固定资产投资、年底就业人数和财政科技投入对数值之间的方差-协方差矩阵。

```
. correlate  lgdp linvest llabor lscientific,covariance
(obs=17)

             |     lgdp  linvest   llabor lscien~c
-------------+------------------------------------
        lgdp |  .497733
     linvest |  .694545  .972207
      llabor |  .035708  .050341  .002839
 lscientific |  .738007   1.0331  .056466  1.17667
```

图 19.24 相关性分析结果图 2

从上述分析结果中，可以看到地区生产总值的对数值的方差是 0.497733，固定资产投资的对数值的方差是 0.972207，年底就业人数的对数值的方差是 0.002839，财政科技投入的对数值的方差是 1.17667，地区生产总值对数值与固定资产投资对数值之间的协方差是 0.694545，地区生产总值对数值与年底就业人数对数值之间的协方差是 0.035708，地区生产总值对数值与财政科技投入对数值之间的协方差是 0.738007，固定资产投资对数值与年底就业人数对数值之间的协方差是 0.050341，固定资产投资对数值与财政科技投入对数值之间的协方差是 1.0331，财政科技投入对数值与年底就业人数对数值之间的协方差是 0.056466。可以发现对变量进行对数变换处理后，变量的方差差距减少了很多，对数变换处理起到了应有的效果。

图 19.25 展示的是变量地区生产总值、固定资产投资、年底就业人数和财政科技投入之间的相关系数矩阵。

```
. correlate  gdp invest labor scientific
(obs=17)

             |      gdp   invest    labor scient~c
-------------+------------------------------------
         gdp |   1.0000
      invest |   0.9978   1.0000
       labor |   0.8541   0.8281   1.0000
  scientific |   0.9877   0.9817   0.8482   1.0000
```

图 19.25　相关性分析结果图 3

从上述分析结果中，可以看到变量地区生产总值、固定资产投资、年底就业人数和财政科技投入之间的相关系数非常高。其中地区生产总值与固定资产投资之间的相关系数是0.9978，地区生产总值与年底就业人数之间的相关系数是 0.8541，地区生产总值与财政科技投入之间的相关系数是 0.9877，固定资产投资与年底就业人数之间的相关系数是 0.8281，固定资产投资与财政科技投入之间的相关系数是 0.9817，财政科技投入与年底就业人数之间的相关系数是 0.8482。各变量之间如此之高的正相关系数在一定程度上说明这几个变量之间很可能存在着一定的联动关系，说明我们的后续分析是很有必要的。

图 19.26 展示的是地区生产总值、固定资产投资、年底就业人数和财政科技投入等变量的对数值之间的相关系数矩阵。

```
. correlate  lgdp linvest llabor lscientific
(obs=17)

             |     lgdp  linvest   llabor lscien~c
-------------+------------------------------------
        lgdp |   1.0000
     linvest |   0.9984   1.0000
      llabor |   0.9500   0.9582   1.0000
 lscientific |   0.9644   0.9659   0.9770   1.0000
```

图 19.26　相关性分析结果图 4

从上述分析结果中，可以看到地区生产总值、固定资产投资、年底就业人数和财政科技投入等变量的对数值之间的相关系数非常高。其中地区生产总值的对数值与固定资产投资的对数值之间的相关系数是 0.9984，地区生产总值的对数值与年底就业人数的对数值之间的相关系数是 0.9500，地区生产总值的对数值与财政科技投入的对数值之间的相关系数是 0.9644，固定资产投资的对数值与年底就业人数的对数值之间的相关系数是 0.9582，固定资产投资的对数值与财政科技投入的对数值之间的相关系数是 0.9659，财政科技投入的对数值与年底就业人数的对数值之间的相关系数是 0.9770。各变量之间如此之高的正相关系数在一定程度上说明这几个变量之间很可能存在着一定的联动关系，说明我们的后续分析是很有必要的。

图 19.27 展示的是变量地区生产总值、固定资产投资、年底就业人数和财政科技投入之间的相关系数矩阵的显著性检验，设定置信水平为 99%。从分析结果中可以看到 4 个变量之间的相关系数非常高，均通过了置信水平为 99%的相关性检验。

图 19.28 展示的是变量地区生产总值、固定资产投资、年底就业人数和财政科技投入之间的相关系数矩阵的显著性检验，设定置信水平为 99%。

```
. pwcorr lgdp linvest llabor lscientific,sidak sig star(0.01)

             |     lgdp  linvest   llabor lscien~c
-------------+------------------------------------
        lgdp |   1.0000 
             |
             |
     linvest |   0.9984*  1.0000 
             |   0.0000
             |
      llabor |   0.9500*  0.9582*  1.0000 
             |   0.0000   0.0000
             |
 lscientific |   0.9644*  0.9659*  0.9770*  1.0000 
             |   0.0000   0.0000   0.0000
             |
```

图 19.27　相关性分析结果图 5

```
. pwcorr gdp invest labor scientific,sidak sig star(0.01)

             |      gdp   invest    labor scient~c
-------------+------------------------------------
         gdp |   1.0000 
             |
             |
      invest |   0.9978*  1.0000 
             |   0.0000
             |
       labor |   0.8541*  0.8281*  1.0000 
             |   0.0001   0.0002
             |
  scientific |   0.9877*  0.9817*  0.8482*  1.0000 
             |   0.0000   0.0000   0.0001
             |
```

图 19.28　相关性分析结果图 6

从上述分析结果中可以看到 4 个变量经对数变换处理之后的相关系数依然非常高，均通过了置信水平为 99%的相关性检验。

19.5　单位根检验

对于时间序列数据而言，数据的平稳性对于模型的构建是非常重要的。如果时间序列数据是不平稳的，可能会导致自回归系数的估计值向左偏向于 0，使传统的 T 检验失效，也有可能会使得两个相互独立的变量出现假相关关系或者回归关系，造成模型结果的失真。单位根检验是判断数据是否平稳的重要方法。只有进行了该步操作才能进行后续深入的分析。

19.5.1　Stata 分析过程

可以发现经过对数变换处理之后的变量要优于原变量，所以在后续的分析中不再包含原变量，只针对对数变换之后的变量进行分析，并得出研究结论。本例我们采用 3 种单位根检验分析方法，分别是 PP 检验、ADF 检验以及 DF-GLS 检验。通过绘制时间序列趋势图可以发现变量地区生产总值、固定资产投资、年底就业人数、财政科技投入的值以及其对数标准化的值都是有明显、稳定的向上增长趋势的，而变量地区生产总值、固定资产投资、年底就业人数和财政科技投入的对数值的一阶差分值是没有明显、稳定的时间趋势的。这些结论将会在单位根检验的操作命令中被用到。

1. PP 检验

操作步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中分别输入如下命令并按键盘上的回车键进行确认。

- pperron　lgdp,trend：本命令旨在对“lgdp”变量运用 PP 检验方法进行单位根检验，以判断该时间序列变量是否平稳。
- pperron　linvest,trend：本命令旨在对“linvest”变量运用 PP 检验方法进行单位根检验，以判断该时间序列变量是否平稳。

- pperron llabor,trend：本命令旨在对“llabor”变量运用PP检验方法进行单位根检验，以判断该时间序列变量是否平稳。
- pperron lscientific,trend：本命令旨在对“lscientific”变量运用PP检验方法进行单位根检验，以判断该时间序列变量是否平稳。
- pperron d.lgdp,notrend：本命令旨在对“d.lgdp”变量运用PP检验方法进行单位根检验，以判断该时间序列变量是否平稳。
- pperron d.lscientific,notrend：本命令旨在对“d.lscientific”变量运用PP检验方法进行单位根检验，以判断该时间序列变量是否平稳。

03 设置完毕后，等待输出结果。

2．ADF检验

操作步骤如下：

01 进入Stata 14.0，打开相关数据文件，弹出主界面。
02 在主界面的“Command”文本框中分别输入如下命令并按键盘上的回车键进行确认。

- dfuller lgdp,trend：本命令旨在对“lgdp”变量运用ADF检验方法进行单位根检验，以判断该时间序列变量是否平稳。
- dfuller linvest,trend：本命令旨在对“linvest”变量运用ADF检验方法进行单位根检验，以判断该时间序列变量是否平稳。
- dfuller llabor,trend：本命令旨在对“llabor”变量运用ADF检验方法进行单位根检验，以判断该时间序列变量是否平稳。
- dfuller lscientific,trend：本命令旨在对“lscientific”变量运用ADF检验方法进行单位根检验，以判断该时间序列变量是否平稳。
- dfuller d.lgdp,notrend：本命令旨在对“d.lgdp”变量运用ADF检验方法进行单位根检验，以判断该时间序列变量是否平稳。
- dfuller d.lscientific,notrend：本命令旨在对“d.lscientific”变量运用ADF检验方法进行单位根检验，以判断该时间序列变量是否平稳。

03 设置完毕后，等待输出结果。

3．DF-GLS检验

操作步骤如下：

01 进入Stata 14.0，打开相关数据文件，弹出主界面。
02 在主界面的“Command”文本框中分别输入如下命令并按键盘上的回车键进行确认。

- dfgls lgdp：本命令旨在对“lgdp”变量运用DF-GLS检验方法进行单位根检验，以判断该时间序列变量是否平稳。
- dfgls linvest：本命令旨在对“linvest”变量运用DF-GLS检验方法进行单位根检验，以判断该时间序列变量是否平稳。
- dfgls llabor：本命令旨在对“llabor”变量运用DF-GLS检验方法进行单位根检验，

以判断该时间序列变量是否平稳。

- dfgls　lscientific：本命令旨在对"lscientific"变量运用 DF-GLS 检验方法进行单位根检验，以判断该时间序列变量是否平稳。

03 设置完毕后，等待输出结果。

19.5.2　结果分析

在 Stata 14.0 主界面的结果窗口可以看到如图 19.29~图 19.44 所示的分析结果。

1．PP 检验的结果

PP 检验的结果如图 19.29~图 19.34 所示。其中图 19.29 展示的是对"lgdp"变量运用 PP 检验方法进行单位根检验的结果。

```
. pperron  lgdp,trend

Phillips-Perron test for unit root                 Number of obs   =        16
                                                   Newey-West lags =         2

                               ——————— Interpolated Dickey-Fuller ———————
                  Test         1% Critical       5% Critical      10% Critical
                Statistic        Value             Value             Value
------------------------------------------------------------------------------
 Z(rho)          -10.950          -22.500           -17.900           -15.600
 Z(t)             -3.015           -4.380            -3.600            -3.240
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.1279
```

图 19.29　单位根检验分析结果图 1

PP 检验的原假设是数据有单位根。从上面的结果中可以看出 P 值（MacKinnon approximate p-value for Z(t)）为 0.1279，接受了有单位根的原假设，这一点也可以通过观察 Z(t)值和 Z(rho)值得到。实际 Z(rho)值为-10.950，在 1%的置信水平（-22.500）、5%的置信水平（-17.900）、10%的置信水平上（-15.600）都无法拒绝原假设。实际 Z(t)值为-3.015，在 1%的置信水平（-4.380）、5%的置信水平（-3.600）、10%的置信水平上（-3.240）都无法拒绝原假设，所以"lgdp"这一变量数据是存在单位根的，需要对其做一阶差分后再继续进行检验。

图 19.30 展示的是对"linvest"变量运用 PP 检验方法进行单位根检验的结果。

```
. pperron  linvest,trend

Phillips-Perron test for unit root                 Number of obs   =        16
                                                   Newey-West lags =         2

                               ——————— Interpolated Dickey-Fuller ———————
                  Test         1% Critical       5% Critical      10% Critical
                Statistic        Value             Value             Value
------------------------------------------------------------------------------
 Z(rho)          -11.641          -22.500           -17.900           -15.600
 Z(t)             -3.965           -4.380            -3.600            -3.240
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0099
```

图 19.30　单位根检验分析结果图 2

PP 检验的原假设是数据有单位根。从上面的结果中可以看出 P 值（MacKinnon approximate p-value for Z(t)）为 0.0099，非常显著地拒绝了有单位根的原假设，这一点也可以通过观察 Z(t)值得到。实际 Z(t)值为-3.965，处在 1%的置信水平（-4.380）与 5%的置信水平（-3.600）之间，

显著地拒绝了有单位根的原假设。

图 19.31 展示的是对“llabor”变量运用 PP 检验方法进行单位根检验的结果。

```
. pperron    llabor,trend

Phillips-Perron test for unit root                 Number of obs    =        16
                                                   Newey-West lags  =         2

                               ---------- Interpolated Dickey-Fuller ----------
                  Test          1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
-------------------------------------------------------------------------------
 Z(rho)         -10.129           -22.500           -17.900           -15.600
 Z(t)           -22.825            -4.380            -3.600            -3.240
-------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0000
```

图 19.31　单位根检验分析结果图 3

PP 检验的原假设是数据有单位根。从上面的结果中可以看出 P 值（MacKinnon approximate p-value for Z(t)）为 0.0000，拒绝了有单位根的原假设，这一点也可以通过观察 Z(t)值得到。实际 Z(t)值为-22.825，在 1%的置信水平（-4.380）、5%的置信水平（-3.600）、10%的置信水平上（-3.240）都拒绝了原假设，所以“llabor”这一变量数据是不存在单位根的。

图 19.32 展示的是对“lscientific”变量运用 PP 检验方法进行单位根检验的结果。

```
. pperron    lscientific,trend

Phillips-Perron test for unit root                 Number of obs    =        16
                                                   Newey-West lags  =         2

                               ---------- Interpolated Dickey-Fuller ----------
                  Test          1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
-------------------------------------------------------------------------------
 Z(rho)          -5.375           -22.500           -17.900           -15.600
 Z(t)            -2.673            -4.380            -3.600            -3.240
-------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.2476
```

图 19.32　单位根检验分析结果图 4

PP 检验的原假设是数据有单位根。从上面的结果中可以看出 P 值（MacKinnon approximate p-value for Z(t)）为 0.2476，显著地接受了有单位根的原假设，这一点也可以通过观察 Z(t)值和 Z(rho)值得到。实际 Z(t)值为-2.673，在 1%的置信水平（-4.380）、5%的置信水平（-3.600）、10%的置信水平上（-3.240）都接受了原假设，实际 Z(rho)值为-5.375，在 1%的置信水平（-22.500）、5%的置信水平（-17.900）、10%的置信水平上（-15.600）都接受了原假设，所以“lscientific”这一变量数据是存在单位根的。

图 19.33 展示的是对“d.lgdp”变量运用 PP 检验方法进行单位根检验的结果。

```
. pperron   d.lgdp,notrend

Phillips-Perron test for unit root                 Number of obs    =        15
                                                   Newey-West lags  =         2

                               ---------- Interpolated Dickey-Fuller ----------
                  Test          1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
-------------------------------------------------------------------------------
 Z(rho)         -10.554           -17.200           -12.500           -10.200
 Z(t)            -4.133            -3.750            -3.000            -2.630
-------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0009
```

图 19.33　单位根检验分析结果图 5

PP 检验的原假设是数据有单位根。从上面的结果中可以看出 P 值（MacKinnon approximate p-value for Z(t)）为 0.0009，显著拒绝了有单位根的原假设，这一点也可以通过观察 Z(t)值得到。实际 Z(t)值为-4.133，在 1%的置信水平（-3.750）、5%的置信水平（-3.000）、10%的置信水平上（-2.630）都拒绝了原假设，所以“d.lgdp”这一变量数据是不存在单位根的。

图 19.34 展示的是对“d.lscientific”变量运用 PP 检验方法进行单位根检验的结果。

```
. pperron   d.lscientific,notrend

Phillips-Perron test for unit root                 Number of obs   =        15
                                                   Newey-West lags =         2

                              ——————— Interpolated Dickey-Fuller ———————
                  Test         1% Critical       5% Critical      10% Critical
               Statistic          Value             Value             Value
──────────────────────────────────────────────────────────────────────────────
 Z(rho)          -14.066          -17.200           -12.500           -10.200
 Z(t)             -3.588           -3.750            -3.000            -2.630
──────────────────────────────────────────────────────────────────────────────
MacKinnon approximate p-value for Z(t) = 0.0060
```

图 19.34　单位根检验分析结果图 6

PP 检验的原假设是数据有单位根。从上面的结果中可以看出 P 值（MacKinnon approximate p-value for Z(t)）为 0.0060，显著地拒绝了有单位根的原假设，这一点也可以通过观察 Z(t)值和 Z(rho)值得到。实际 Z(t)值为-3.588，处于 1%的置信水平（-3.750）与 5%的置信水平（-3.000）之间，拒绝了原假设。实际 Z(rho)值为-14.066，处于 1%的置信水平（-17.200）与 5%的置信水平（-12.500）之间，拒绝了原假设，所以“d.lscientific”这一变量数据是不存在单位根的。

2．ADF 检验的结果

ADF 检验的结果如图 19.35~图 19.40 所示。其中图 19.35 展示的是对“lgdp”变量运用 ADF 检验方法进行单位根检验的结果。

```
. dfuller   lgdp,trend

Dickey-Fuller test for unit root                   Number of obs   =        16

                              ——————— Interpolated Dickey-Fuller ———————
                  Test         1% Critical       5% Critical      10% Critical
               Statistic          Value             Value             Value
──────────────────────────────────────────────────────────────────────────────
 Z(t)             -3.066           -4.380            -3.600            -3.240
──────────────────────────────────────────────────────────────────────────────
MacKinnon approximate p-value for Z(t) = 0.1145
```

图 19.35　单位根检验分析结果图 7

ADF 检验的原假设是数据有单位根。从上面的结果中可以看出 P 值（MacKinnon approximate p-value for Z(t)）为 0.1145，接受了有单位根的原假设，这一点也可以通过观察 Z(t)值得到验证。实际 Z(t)值为-3.066，在 1%的置信水平（-4.380）、5%的置信水平（-3.600）、10%的置信水平上（-3.240）都无法拒绝原假设，所以“lgdp”这一变量数据是存在单位根的，需要对其做一阶差分后再继续进行检验。

图 19.36 展示的是对“linvest”变量运用 ADF 检验方法进行单位根检验的结果。

```
. dfuller  linvest,trend

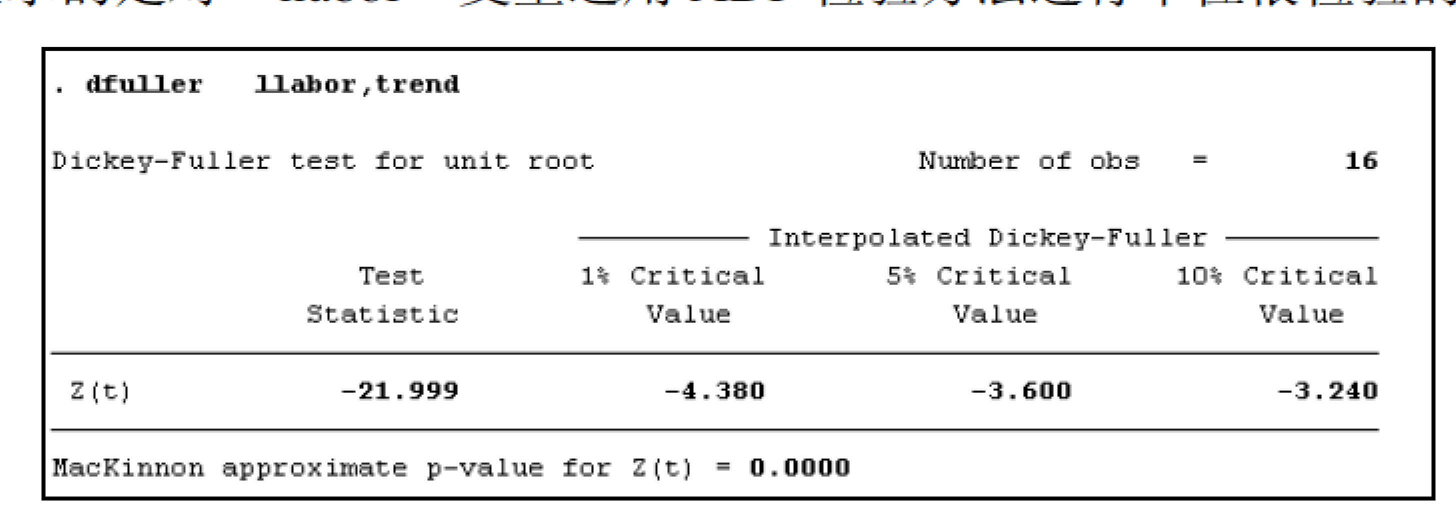
Dickey-Fuller test for unit root                   Number of obs   =        16

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -4.466            -4.380            -3.600            -3.240
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0017
```

图 19.36　单位根检验分析结果图 8

ADF 检验的原假设是数据有单位根。从上面的结果中可以看出 P 值（MacKinnon approximate p-value for Z(t)）为 0.0017，非常显著地拒绝了有单位根的原假设，这一点也可以通过观察 Z(t)值得到。实际 Z(t)值为-4.466，在 1%的置信水平(-4.380)、5%的置信水平(-3.600)、10%的置信水平（-3.240）上都显著拒绝了有单位根的原假设，所以“linvest”这一变量数据是不存在单位根的。

图 19.37 展示的是对“llabor”变量运用 ADF 检验方法进行单位根检验的结果。



```
. dfuller  llabor,trend

Dickey-Fuller test for unit root                   Number of obs   =        16

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)            -21.999            -4.380            -3.600            -3.240
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0000
```

图 19.37　单位根检验分析结果图 9

ADF 检验的原假设是数据有单位根。从上面的结果中可以看出 P 值（MacKinnon approximate p-value for Z(t)）为 0.0000，拒绝了有单位根的原假设，这一点也可以通过观察 Z(t)值得到验证。实际 Z(t)值为-21.999，在 1%的置信水平（-4.380）、5%的置信水平（-3.600）、10%的置信水平（-3.240）上都拒绝了原假设，所以“llabor”这一变量数据是不存在单位根的。

图 19.38 展示的是对“lscientific”变量运用 ADF 检验方法进行单位根检验的结果。

```
. dfuller   lscientific,trend

Dickey-Fuller test for unit root                   Number of obs   =        16

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -2.576            -4.380            -3.600            -3.240
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.2911
```

图 19.38　单位根检验分析结果图 10

ADF 检验的原假设是数据有单位根。从上面的结果中可以看出 P 值（MacKinnon approximate p-value for Z(t)）为 0.2911，显著地接受了有单位根的原假设，这一点也可以通过观察 Z(t)值得到。实际 Z(t)值为-2.576，在 1%的置信水平（-4.380）、5%的置信水平（-3.600）、10%的置信水平上（-3.240）都接受了原假设，所以“lscientific”这一变量数据是存在单位根的。

图 19.39 展示的是对“d.lgdp”变量运用 ADF 检验方法进行单位根检验的结果。

```
. dfuller  d.lgdp,notrend

Dickey-Fuller test for unit root                   Number of obs   =        15

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -3.990            -3.750            -3.000            -2.630
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0015
```

图 19.39　单位根检验分析结果图 11

ADF 检验的原假设是数据有单位根。从上面的结果中可以看出 P 值（MacKinnon approximate p-value for Z(t)）为 0.0015，显著拒绝了有单位根的原假设，这一点也可以通过观察 Z(t)值得到。实际 Z(t)值为-3.990，在 1%的置信水平（-3.750）、5%的置信水平（-3.000）、10%的置信水平（-2.630）上都拒绝了原假设，所以“d.lgdp”这一变量数据是不存在单位根的。

图 19.40 展示的是对“d.lscientific”变量运用 ADF 检验方法进行单位根检验的结果。

```
. dfuller    d.lscientific,notrend

Dickey-Fuller test for unit root                   Number of obs   =        15

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -3.590            -3.750            -3.000            -2.630
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0060
```

图 19.40　单位根检验分析结果图 12

ADF 检验的原假设是数据有单位根。从上面的结果中可以看出 P 值（MacKinnon approximate p-value for Z(t)）为 0.0060，显著地拒绝了有单位根的原假设，这一点也可以通过观察 Z(t)值得到。实际 Z(t)值为-3.590，处于 1%的置信水平（-3.750）与 5%的置信水平（-3.000）之间，拒绝了原假设，所以“d.lscientific”这一变量数据是不存在单位根的。

3．DF-GLS 检验的结果

DF-GLS 检验的结果如图 19.41~图 19.44 所示。其中图 19.41 展示的是“lgdp”变量的 DF-GLS 检验结果。

```
. dfgls  lgdp

DF-GLS for lgdp                                          Number of obs =     9
Maxlag = 7 chosen by Schwert criterion

               DF-GLS tau      1% Critical       5% Critical      10% Critical
  [lags]     Test Statistic        Value             Value             Value
------------------------------------------------------------------------------
    7            -0.713           -3.770            -7.782            -5.617
    6            -0.782           -3.770            -5.328            -3.779
    5            -1.533           -3.770            -3.828            -2.701
    4            -3.408           -3.770            -3.080            -2.217
    3            -1.679           -3.770            -2.882            -2.159
    2            -1.951           -3.770            -3.032            -2.361
    1            -0.928           -3.770            -3.326            -2.655

Opt Lag (Ng-Perron seq t) =  4 with RMSE   .010138
Min SC    = -7.962249 at lag  4 with RMSE   .010138
Min MAIC = -7.116174 at lag  1 with RMSE  .0219052
```

图 19.41　单位根检验分析结果图 13

DF-GLS 检验的原假设是数据有单位根。从上面的结果中可以看出根据信息准则确定的最优滞后阶数为 4 阶（Opt Lag (Ng-Perron seq t) = 4with RMSE 0.010138），在该阶数下 DF-GLS 统计量的值是-3.408，处于 1%的置信水平（-3.770）与 5%的置信水平（-3.080）之间，拒绝了有单位根的原假设，所以“lgdp”变量数据是不存在单位根的。这一点显然与我们前面两种方法的检验结果不一致，但这也是正常情况，事实上我们选择多种检验方法对数据进行单位根检验的初衷就是综合各种检验方法的检验结果做出恰当的判断。

图 19.42 展示的是“linvest”变量的 DF-GLS 检验结果。

```
. dfgls   linvest

DF-GLS for linvest                                       Number of obs =     9
Maxlag = 7 chosen by Schwert criterion

               DF-GLS tau      1% Critical       5% Critical      10% Critical
  [lags]     Test Statistic       Value             Value             Value
------------------------------------------------------------------------------
    7            -1.528           -3.770           -7.782            -5.617
    6            -2.977           -3.770           -5.328            -3.779
    5            -4.065           -3.770           -3.828            -2.701
    4            -4.829           -3.770           -3.080            -2.217
    3            -7.066           -3.770           -2.882            -2.159
    2            -2.711           -3.770           -3.032            -2.361
    1            -2.960           -3.770           -3.326            -2.655

Opt Lag (Ng-Perron seq t) =  3 with RMSE  .0113834
Min SC    = -7.974657 at lag  3 with RMSE  .0113834
Min MAIC = -3.247334 at lag  1 with RMSE  .0288839
```

图 19.42　单位根检验分析结果图 14

DF-GLS 检验的原假设是数据有单位根。从上面的结果中可以看出根据信息准则确定的最优滞后阶数为 3 阶（Opt Lag (Ng-Perron seq t) = 3with RMSE.0113834），在该阶数下 DF-GLS 统计量的值是-7.066，在 1%的置信水平（-3.770）、5%的置信水平（-2.882）、10%的置信水平（-2.159）上都显著拒绝了有单位根的原假设，所以“linvest”变量数据是不存在单位根的。

图 19.43 展示的是对“llabor”变量运用 DF-GLS 检验方法进行单位根检验的结果。

```
. dfgls    llabor

DF-GLS for llabor                                        Number of obs =     9
Maxlag = 7 chosen by Schwert criterion

               DF-GLS tau      1% Critical       5% Critical      10% Critical
  [lags]     Test Statistic       Value             Value             Value
------------------------------------------------------------------------------
    7            -0.997           -3.770           -7.782            -5.617
    6            -0.955           -3.770           -5.328            -3.779
    5            -2.049           -3.770           -3.828            -2.701
    4            -1.146           -3.770           -3.080            -2.217
    3            -1.002           -3.770           -2.882            -2.159
    2            -1.004           -3.770           -3.032            -2.361
    1            -0.510           -3.770           -3.326            -2.655

Opt Lag (Ng-Perron seq t) = 0 [use maxlag(0)]
Min SC    = -11.10706 at lag  7 with RMSE  .0014589
Min MAIC =   -10.9429 at lag  1 with RMSE  .0035652
```

图 19.43　单位根检验分析结果图 15

DF-GLS 检验的原假设是数据有单位根。从上面的结果中可以看出根据信息准则确定的最优滞后阶数为 0 阶（Opt Lag (Ng-Perron seq t) = 0 [use maxlag(0)]）。但是结果中并没有 0 阶的体现，我们可以观测根据 MAIC 信息准则确定的 1 阶（Min MAIC = -10.9429 at lag 1 with

RMSE 0.0035652）来判定结果，在 1 阶的时候，接受了原假设，变量数据是存在单位根的。这一点显然与我们前面两种方法的检验结果不一致，但这也是正常情况，事实上我们选择多种检验方法对数据进行单位根检验的初衷就是综合各种检验方法的检验结果做出恰当的判断。

图 19.44 展示的是对“lscientific”变量运用 DF-GLS 检验方法进行单位根检验的结果。

```
. dfgls   lscientific

DF-GLS for lscientific                                   Number of obs =     9
Maxlag = 7 chosen by Schwert criterion

               DF-GLS tau      1% Critical       5% Critical       10% Critical
  [lags]     Test Statistic       Value             Value              Value
------------------------------------------------------------------------------
    7           -14.270          -3.770            -7.782             -5.617
    6            -1.646          -3.770            -5.328             -3.779
    5            -2.445          -3.770            -3.828             -2.701
    4            -1.959          -3.770            -3.080             -2.217
    3            -3.042          -3.770            -2.882             -2.159
    2            -2.376          -3.770            -3.032             -2.361
    1            -2.584          -3.770            -3.326             -2.655

Opt Lag (Ng-Perron seq t) =  7 with RMSE   .008676
Min SC    = -7.541297 at lag  7 with RMSE   .008676
Min MAIC = -.9577767 at lag  1 with RMSE  .1453201
```

图 19.44 单位根检验分析结果图 16

DF-GLS 检验的原假设是数据有单位根。从上面的结果中可以看出根据信息准则确定的最优滞后阶数为 7 阶（Opt Lag (Ng-Perron seq t) = 7 with RMSE 0.008676），在该阶数下 DF-GLS 统计量的值是-14.270，拒绝了原假设，不存在单位根。这一点显然与我们前面两种方法的检验结果不一致，但这也是正常情况，事实上我们选择多种检验方法对数据进行单位根检验的初衷就是综合各种检验方法的检验结果做出恰当的判断。

根据以上的分析，综合考虑三种检验方法的检验结果，我们可以比较有把握地得出以下结论，即认为变量地区生产总值的对数值、财政科技投入的对数值是存在单位根的，变量固定资产投资的对数值、年底就业人数的对数值、地区生产总值的对数值的一阶差分值、财政科技投入的对数值的一阶差分值是不存在单位根的。在该结论的基础上，我们将进入下一节的协整检验分析过程。

19.6 协整检验

在时间序列数据不平稳的情况下，构建出合理模型的重要方法就是进行协整检验并构建合理模型。协整的思想就是把存在一阶单整的变量放在一起进行分析，通过这些变量进行线性组合，从而消除他们的随机趋势，得到其长期联动趋势。

19.6.1 Stata 分析过程

本例我们采用迹检验协整检验分析方法。在前面几节中，我们通过绘制时间序列趋势图发现变量地区生产总值、固定资产投资、年底就业人数和财政科技投入的值以及其对数标准化

的值都是有明显、稳定的向上增长趋势的，而变量地区生产总值、固定资产投资、年底就业人数和财政科技投入的对数值的一阶差分值是没有明显、稳定的时间趋势的。通过 PP 检验、ADF 检验以及 DF-GLS 检验等单位根检验发现变量地区生产总值的对数值、财政科技投入的对数值是存在单位根的，变量固定资产投资的对数值、年底就业人数的对数值、地区生产总值的对数值的一阶差分值、财政科技投入的对数值的一阶差分值是不存在单位根的。变量地区生产总值的对数值、财政科技投入的对数值是一阶单整的。这些结论将会在协整检验的操作命令中被用到。

本例中，因为仅有变量地区生产总值的对数值、财政科技投入的对数值是非平稳且一阶单整的，所以只研究这两个变量之间的长期均衡关系是否存在。迹检验的操作步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的"Command"文本框中分别输入如下命令并按键盘上的回车键进行确认。

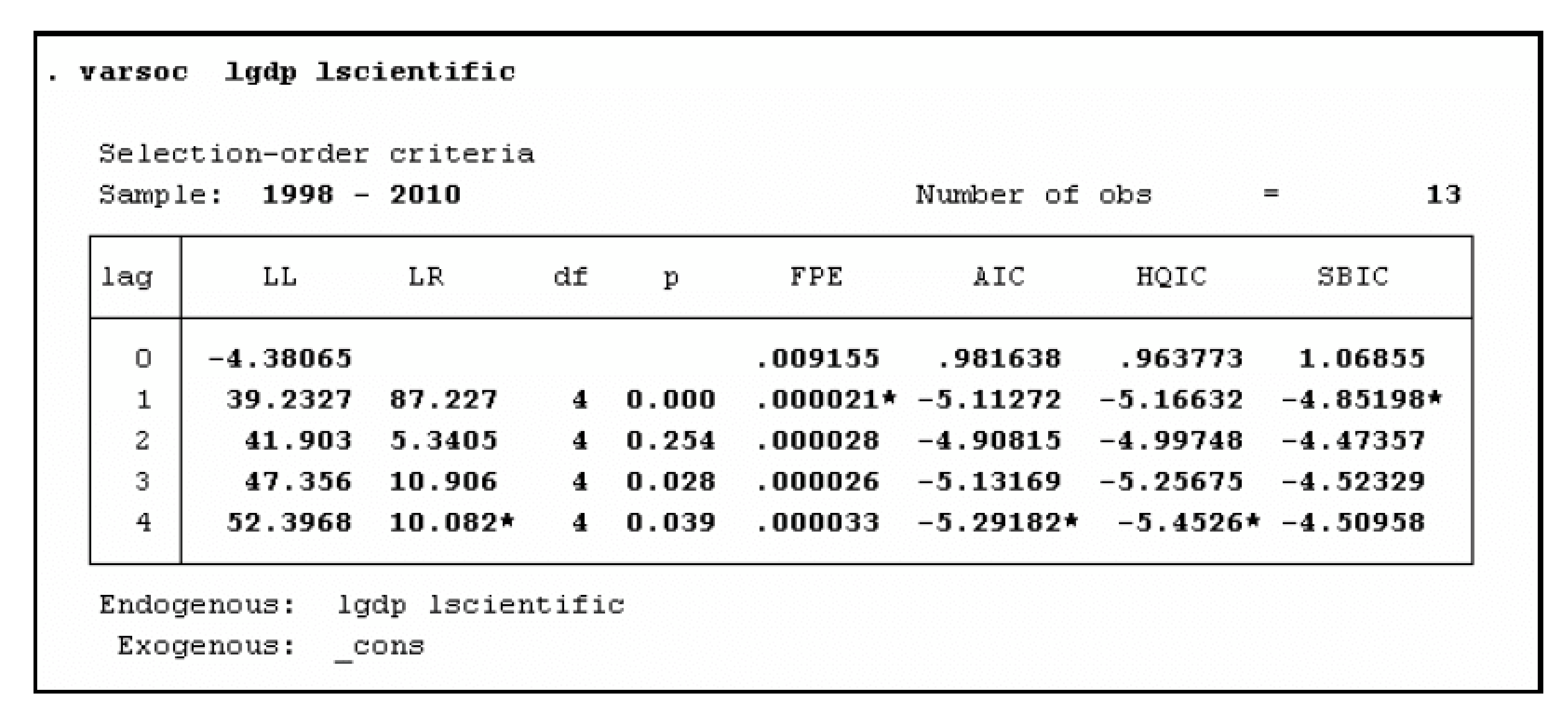
- varsoc lgdp lscientific：本命令的主要目的是根据信息准则确定变量的滞后阶数。
- vecrank lgdp lscientific,lags(4)：本命令的主要目的是在确定滞后阶数的基础上，确定协整秩。
- vecrank lgdp lscientific,lags(1)：本命令的主要目的同样是在确定滞后阶数的基础上，确定协整秩。

03 设置完毕后，等待输出结果。

19.6.2 结果分析

目前国际上公认的比较合理的信息准则有很多种，所以研究者在选取滞后阶数时要适当加入自己的判断。在确定滞后阶数后，我们要确定协整秩，协整秩代表着协整关系的个数。变量之间往往会存在多个长期均衡关系，所以协整秩并不必然等于 1。在确定协整秩后，我们就可以构建相应的模型，写出协整方程了。

在 Stata 14.0 主界面的结果窗口可以看到如图 19.45~图 19.47 所示的分析结果。



```
. varsoc  lgdp lscientific

  Selection-order criteria
  Sample:  1998 - 2010                          Number of obs     =        13
```

lag	LL	LR	df	p	FPE	AIC	HQIC	SBIC
0	-4.38065				.009155	.981638	.963773	1.06855
1	39.2327	87.227	4	0.000	.000021*	-5.11272	-5.16632	-4.85198*
2	41.903	5.3405	4	0.254	.000028	-4.90815	-4.99748	-4.47357
3	47.356	10.906	4	0.028	.000026	-5.13169	-5.25675	-4.52329
4	52.3968	10.082*	4	0.039	.000033	-5.29182*	-5.4526*	-4.50958

```
  Endogenous:  lgdp lscientific
   Exogenous:  _cons
```

图 19.45 协整检验分析结果图 1

图 19.45 给出了根据信息准则确定的变量滞后阶数分析结果。最左列的 lag 表示的是滞后阶数，LL、LR 两列表示的是统计量，df 表示的是自由度，p 值表示的是对应滞后阶数下模型

的显著性，FPE、AIC、HQIC、SBIC 代表的是 4 种信息准则，其中值越小越好，越应该选用，这一点也可以通过观察“*”号来验证，带“*”号说明在本信息准则下的最优滞后阶数。最下面两行文字说明的是模型中的外生变量和内生变量，本例中，外生变量包括 lgdp、lscientific（Endogenous: lgdp lscientific），内生变量包括常数项（Exogenous: _cons）。

综上所述，可以看出选取滞后阶数为 1 阶或者 4 阶是比较合适的，下面我们分别来判断一下两种滞后阶数下协整秩的具体情况。

当滞后阶数为 4 时，结果如图 19.46 所示。

```
. vecrank lgdp lscientific,lags(4)

                      Johansen tests for cointegration
Trend: constant                                         Number of obs =      13
Sample:  1998 - 2010                                             Lags =       4
-------------------------------------------------------------------------------
                                                             5%
maximum                                          trace    critical
  rank    parms       LL       eigenvalue  statistic    value
    0      14     41.042331         .       22.7089     15.41
    1      17     52.381008     0.82525      0.0316*     3.76
    2      18     52.396801     0.00243
-------------------------------------------------------------------------------
```

图 19.46 协整检验分析结果图 2

图 19.46 展示的是根据前面确定的滞后阶数确定协整秩的结果。分析本结果最直接的方式就是找到带“*”号的迹统计量（trace statistic），本例中该值为 0.0316，对应的协整秩为 1，这说明本例中地区生产总值的对数值、财政科技投入的对数值两个变量存在一个协整关系。

当滞后阶数为 1 时，结果如图 19.47 所示。

```
. vecrank lgdp lscientific,lags(1)

                      Johansen tests for cointegration
Trend: constant                                         Number of obs =      16
Sample:  1995 - 2010                                             Lags =       1
-------------------------------------------------------------------------------
                                                             5%
maximum                                          trace    critical
  rank    parms       LL       eigenvalue  statistic    value
    0       2     34.306019         .       25.7607     15.41
    1       5     46.951119     0.79416      0.4705*     3.76
    2       6     47.186348     0.02898
-------------------------------------------------------------------------------
```

图 19.47 协整检验分析结果图 3

图 19.47 展示的是根据前面确定的滞后阶数确定协整秩的结果。分析本结果最直接的方式就是找到带“*”号的迹统计量（trace statistic），本例中该值为 0.4705，对应的协整秩为 1，这说明本例中地区生产总值的对数值、财政科技投入的对数值两个变量存在一个协整关系。

至此，协整检验已毕。我们发现两种滞后阶数得到的结论是一致的。对于迹检验而言，同样可以构建出相应的模型来描述这种长期协整关系。这一点将在后续的“建立模型”一节中进行详细说明。

19.7 格兰杰因果关系检验

协整关系表示的仅仅是变量之间的某种长期联动关系，与因果关系是毫无关联的，例如本例中虽然地区生产总值的对数值、财政科技投入的对数值两个变量之间存在协整关系，但是究竟是地区生产总值的对数值影响了财政科技投入的对数值，还是财政科技投入的对数值影响了地区生产总值的对数值，亦或是它们相互影响？如果要探究变量之间的因果关系，就需要用到格兰杰因果关系检验。

19.7.1 Stata 分析过程

在前面几节中，通过单位根检验发现地区生产总值的对数值、财政科技投入的对数值两个变量是一阶单整的，所以我们在进行格兰杰因果关系检验时选择的变量是：地区生产总值的对数值、财政科技投入的对数值。

格兰杰因果关系检验的操作步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中分别输入如下命令并按键盘上的回车键进行确认。

- reg lgdp l.lgdp l.lscientific：本命令旨在以地区生产总值的对数值为因变量，以地区生产总值的对数值的滞后一期值、财政科技投入的对数值的滞后一期值为自变量，进行最小二乘回归分析。
- test l.lscientific：本命令旨在检验财政科技投入的对数值的滞后一期值这一变量的系数是否显著。
- reg lscientific l.lscientific l.lgdp：本命令旨在以财政科技投入的对数值为因变量，以财政科技投入的对数值的滞后一期值、地区生产总值的对数值的滞后一期值为自变量，进行最小二乘回归分析。
- test l.lgdp：本命令旨在检验地区生产总值的对数值的滞后一期值这一变量的系数是否显著。

03 设置完毕后，等待输出结果。

19.7.2 结果分析

在 Stata 14.0 主界面的结果窗口可以看到如图 19.48~图 19.51 所示的分析结果。

图 19.48 和图 19.49 展示的是财政科技投入是否是地区生产总值的格兰杰因的检验结果。通过观察分析结果可以看出 l.lscientific 的系数值是非常显著的。具体体现在其 t 值、F 值以及 P 值上，关于这一结果的详细解读方法前面章节中多有提及，限于篇幅此处不再赘述，所以我们可以比较有把握地得出结论，财政科技投入是地区生产总值的格兰杰因。

```
. reg   lgdp l.lgdp l.lscientific

      Source |       SS       df       MS              Number of obs =      16
-------------+------------------------------           F(  2,     13) = 5284.71
       Model |  6.39016733     2  3.19508366           Prob > F      =  0.0000
    Residual |  .00785967     13  .00060459            R-squared     =  0.9988
-------------+------------------------------           Adj R-squared =  0.9986
       Total |   6.398027     15  .426535133           Root MSE      =  .02459

------------------------------------------------------------------------------
        lgdp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lgdp |
         L1. |   1.119264   .0346644    32.29   0.000     1.044376    1.194152
             |
 lscientific |
         L1. |  -.0887465   .0218954    -4.05   0.001    -.1360485   -.0414444
             |
       _cons |    .145054   .0734631     1.97   0.070    -.0136534    .3037614
------------------------------------------------------------------------------
```

图 19.48　格兰杰因果关系检验分析结果图 1

```
. test  l.lscientific

 ( 1)  L.lscientific = 0

       F(  1,    13) =   16.43
            Prob > F =   0.0014
```

图 19.49　格兰杰因果关系检验分析结果图 2

图 19.50 和图 19.51 展示的是地区生产总值是否是财政科技投入的格兰杰因的检验结果。通过观察分析结果可以看出 l.lgdp 的系数值是非常显著的。具体体现在其 t 值、F 值以及 P 值上，关于这一结果的详细解读方法前面章节中多有提及，限于篇幅此处不再赘述，综上所述，我们可以比较有把握地认为地区生产总值与财政科技投入互为格兰杰因。

```
. reg  lscientific l.lscientific l.lgdp

      Source |       SS       df       MS              Number of obs =      16
-------------+------------------------------           F(  2,     13) =  241.05
       Model |  12.8919269     2  6.44596344           Prob > F      =  0.0000
    Residual |  .347635858    13  .02674122            R-squared     =  0.9737
-------------+------------------------------           Adj R-squared =  0.9697
       Total |  13.2395627    15  .882637516           Root MSE      =  .16353

------------------------------------------------------------------------------
 lscientific |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 lscientific |
         L1. |   .5568393    .145617     3.82   0.002     .2422528    .8714257
             |
        lgdp |
         L1. |   .5304502   .2305388     2.30   0.039     .0324013    1.028499
             |
       _cons |   .6854741    .488573     1.40   0.184    -.3700237    1.740972
------------------------------------------------------------------------------
```

图 19.50　格兰杰因果关系检验分析结果图 3

```
. test l.lgdp

 ( 1)  L.lgdp = 0

       F(  1,    13) =    5.29
            Prob > F =    0.0386

.
```

图 19.51　格兰杰因果关系检验分析结果图 4

19.8 建立模型

本节将执行最后的步骤，即根据前面得出的一系列结论建立相应的数据模型。建立模型的步骤如下。

1．建立模型方程

根据前面几节的分析，构建如下所示的模型方程：

$$d.lgdp = \alpha\ linvest + \beta llabor + \gamma d.lscientific + lnAt + \mu$$

其中，gdp、invest、labor、scientific 分别表示地区生产总值、固定资产投资、年底就业人数和财政科技投入。α、β 和 γ 分别表示固定资产投资、年底就业人数和财政科技投入的产出弹性，lnAt 为常数项，而 μ 是随机误差项。

2．估计整体方程

在主界面的“Command”文本框中输入命令：

```
reg d.lgdp linvest llabor d.lscientific
```

并按键盘上的回车键进行确认，即可出现如图 19.52 所示的模型整体方程估计结果。

```
. reg d.lgdp linvest llabor d.lscientific

      Source |       SS           df       MS      Number of obs   =        16
-------------+----------------------------------   F(  3,     12)  =     11.13
       Model |  .014292852         3  .004764284   Prob > F        =    0.0009
    Residual |  .005136272        12  .000428023   R-squared       =    0.7356
-------------+----------------------------------   Adj R-squared   =    0.6696
       Total |  .019429123        15  .001295275   Root MSE        =    .02069

------------------------------------------------------------------------------
      D.lgdp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     linvest |   .1716194   .0349844     4.91   0.000      .095395    .2478437
      llabor |  -4.084321   .7993328    -5.11   0.000    -5.825917   -2.342725
             |
 lscientific |
         D1. |  -.0032197   .0291349    -0.11   0.914    -.0666992    .0602599
             |
       _cons |   23.03791   4.477924     5.14   0.000     13.28136    32.79447
------------------------------------------------------------------------------
```

图 19.52 建立模型分析结果图 1

从上述分析结果中可以看到共有 47 个样本参与了分析。模型的 F 值(3, 12) = 11.13，P 值（Prob > F）= 0.0009，说明模型整体上还是可以接受的。模型的可决系数（R-squared）为 0.7356，模型修正的可决系数（Adj R-squared）为 0.6696，说明模型解释能力还是比较不错的。

模型的回归方程是：

d.lgdp =0.1716194* linvest -4.084321 * llabor -0.0032197*d1.lscientific + 23.03791

变量 linvest 的系数标准误是 0.0349844，t 值为 4.91，P 值为 0.000，系数是非常显著的，95%的置信区间为[0.095395，0.2478437]。变量 llabor 的系数标准误是 0.7993328，t 值为-5.11，P 值为 0.000，系数也是非常显著的，95%的置信区间为[-5.825917 , -2.342725]。变量 d1.lscientific 的系数标准误是 0.0291349，t 值为-0.11，P 值为 0.914，系数是非常不显著的，95%的置信区间为[-0.0666992, 0.0602599]。常数项的系数标准误是 4.477924，t 值为 5.14，P 值为 0.000，系

数也是非常显著的，95%的置信区间为[13.28136 , 32.79447]。

需要特别解释的是济南市的经济持续增长是一种事实，而且根据经济增长理论，资本（固定资产投资）、劳动力（年底就业人数）、科技投入（财政科技投入）对经济增长都是有促进作用的，所以 d.lgdp 反映的是经济增长的差额，或者说经济增长的速度。从该模型方程中可以得到很多信息：

- 首先，固定资产投资的系数为正而且非常显著，这说明济南市的固定资产投资对地区生产总值的变化是具有显著的正向作用的，在一定程度上说明了粗放的固定资产投资仍是济南市的重要经济增长动力，固定投资越多，经济增长越快。
- 其次，年底就业人数的系数为负而且非常显著，这说明济南市的年底就业人数对地区生产总值的变化是具有显著的负向作用的，在一定程度上说明了济南市的就业市场已经饱和，过多的就业人口反而会降低经济运行效率，减缓经济增长的速度。而科技投入对地区生产总值的影响变化关系在短期内是不够显著的，说明济南市对科技的投入在短期内的效果不明显，或者说科技投入不能立竿见影，并没有成为济南市经济发展的近期动力。

在主界面的“Command”文本框中输入命令：

```
vec lgdp lscientific,lags(1) rank(1)
```

并按键盘上的回车键进行确认，即可出现如图 19.53 所示的地区生产总值与财政科技投入的长期均衡关系模型方程估计结果。

```
Vector error-correction model

Sample:  1995 - 2010                            No. of obs      =        16
                                                AIC             =  -5.24389
Log likelihood =  46.95112                      HQIC            = -5.231527
Det(Sigma_ml)  =  9.69e-06                      SBIC            = -5.002456

Equation           Parms      RMSE     R-sq      chi2     P>chi2
----------------------------------------------------------------
D_lgdp               2     .023913   0.9781   624.8607   0.0000
D_lscientific        2     .159261   0.7808   49.86935   0.0000
----------------------------------------------------------------

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
D_lgdp       |
        _ce1 |
         L1. |   .1072967   .0240066     4.47   0.000     .0602446    .1543488
             |
       _cons |   .1028637   .0115501     8.91   0.000      .080226    .1255015
-------------+----------------------------------------------------------------
D_lscientific|
        _ce1 |
         L1. |     .61595   .1598825     3.85   0.000     .3025861    .9293139
             |
       _cons |  -.0179186    .076923    -0.23   0.816    -.1686848    .1328477
------------------------------------------------------------------------------

Cointegrating equations

Equation           Parms    chi2     P>chi2
-------------------------------------------
_ce1                 1   545.5472   0.0000
-------------------------------------------

Identification:  beta is exactly identified

                 Johansen normalization restriction imposed
------------------------------------------------------------------------------
        beta |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_ce1         |
        lgdp |          1          .        .       .            .           .
 lscientific |  -.7791581   .0333587   -23.36   0.000      -.84454   -.7137762
       _cons |   .7267954          .        .       .            .           .
------------------------------------------------------------------------------
```

图 19.53 建立模型分析结果图 2

观察分析结果得到的协整方程为：

e=lgdp-0.7791581*lscientific+0.7267954

该方程反映的是地区生产总值与财政科技投入的长期均衡关系。令 e=0，将模型进行变形可得：

lgdp=0.7791581*lscientific-0.7267954

这个方程说明的是济南市科技投入对地方生产总值的长期作用是正向的，而且非常显著（观察到 lscientific 变量系数的显著性 P 值为 0.000），效果非常明显，能达到近 80%。

根据格兰杰因果关系检验的结果，地区生产总值与财政科技投入的长期均衡关系模型方程为：

d.lgdp=0.1072967*l.e+0.1028637

其中：

e=lgdp-0.7791581*lscientific+0.7267954

d.lgdp=0.1072967*（l.lgdp-0.7791581*l.lscientific+0.7267954）+0.1028637

l.lscientific 前面的系数为负值，说明上期科技投入偏多时，会引起本期地区生产总值的减少。这在一定程度上验证了前面得出的结论，科技投入虽然从长期来看对济南市经济增长贡献将会非常大，但在现阶段达不到预期效果，所以，综上所述，我们可以比较有把握地说，济南市目前的经济增长还是比较粗放的，距离集约型增长模式还有一段比较长的路要走。

19.9　研究结论

经过前面的研究之后，可以比较有把握地得出以下研究结论：

- 变量地区生产总值、固定资产投资、年底就业人数和财政科技投入的值以及其对数标准化的值都是有明显、稳定的向上增长趋势的，而变量地区生产总值、固定资产投资、年底就业人数和财政科技投入的对数值的一阶差分值是没有明显、稳定的时间趋势的。
- 地区生产总值、固定资产投资、年底就业人数和财政科技投入等变量之间的相关系数以及其对数值之间的相关系数都非常高，而且相关关系非常显著。
- 变量地区生产总值的对数值、财政科技投入的对数值是存在单位根的，变量固定资产投资的对数值、年底就业人数的对数值、地区生产总值的对数值的一阶差分值、财政科技投入的对数值的一阶差分值是不存在单位根的。
- 地区生产总值的对数值、财政科技投入的对数值两个变量存在一个协整关系。
- 地区生产总值与财政科技投入互为格兰杰因。
- 济南市的固定资产投资对地区生产总值的变化是具有显著的正向作用的，在一定程度上说明了粗放的固定资产投资仍是济南市的重要经济增长动力，固定投资越多，经济增长越快。济南市的年底就业人数对地区生产总值的变化是具有显著的负向作用的，济南市的就业市场已经饱和，过多的就业人口反而会降低经济运行效率，减缓经济增

长的速度。

- 济南市科技投入对地方生产总值的长期作用是正向的，而且非常显著，能达到近 80%。科技投入虽然从长期来看对济南市经济增长贡献将会非常大，但在现阶段科技投入对地区生产总值的影响变化关系不够显著，或者说科技投入并没有成为济南市经济发展的近期动力。济南市目前的经济增长还是比较粗放的，距离集约型增长模式还有一段比较长的路要走。

19.10　本章习题

表 19.2 给出了某企业经营利润、固定资产投资、员工薪酬和科技研发投入的有关数据，试使用描述性分析、时间序列趋势图分析、相关性检验、单位根检验、协整检验、格兰杰因果关系检验等方法研究数据特征并对变量间的关系进行分析，最后建立相应的方程模型描述变量之间的联动关系。

表 19.2　习题 19 数据

年份	利润/万元	固定资产投资/万元	员工薪酬/万元	科技研发投入/万元
1996	494.76	97.356	40.432	19.045 6
1997	629.781 6	149.970 8	43.118 6	30.683 1
1998	757.105 16	200.564	44.195 9	48.332 2
1999	884.428 72	241.049 2	44.874 2	81.435 9
2000	1 011.752 28	293.517 7	45.432 8	102.237 1
2001	1 139.075 84	359.658 6	45.818 5	168.245
2002	1 266.399 4	406.913 5	46.204 2	173.259 1
2003	1 417.992 8	457.719 5	46.563 3	248.164 7
2004	1 597.103 9	538.237 7	46.909 1	268.447 2
2005	1 815.888 9	671.503 7	47.254 9	194.047
2006	2 153.097 1	866.229	47.680 5	269.338 3
2007	2 455.552 4	1 139.81	47.88	297.693 9
2008	2 874.834 9	1 352.304 1	48.119 4	366.242 1
2009	3 325.186 2	1 531.761	48.451 9	538.862 8
2010	3 999.004 1	1 882.388 9	48.864 2	599.324 6
2011	4 443.410 3	2 201.642 1	49.515 9	699.912 5
2012	5 201.004 9	2 643.295 2	49.702 1	826.435 4

第 20 章 Stata 在原油与黄金价格联动关系研究中的应用

黄金和原油同属于大宗商品，都是衡量宏观经济状况的重要指标，在人类社会发展的过程中都扮演着重要的角色。黄金是公认的硬通货，而原油自工业革命以来，成为现代社会的血液。黄金和原油的价格问题也一直深受社会各界的密切关注，成为专家学者研究分析的热点课题之一。无数国内外学者的研究发现黄金价格和原油价格之间是存在一定联动关系的，它们的价格变化存在着一定的内在规律。当然学者们得出的研究结论并不是完全一致的，有的学者认为黄金和原油存在着一定的正向变动关系，当国际原油价格上扬时，黄金价格常常也随之走高；反之，当油价下跌时，金价亦随之踏空。也有学者持有恰好相反的意见，他们认为原油和黄金在保值增值方面是一种逆向变动关系，油价和金价的变动关系是相反的。还有的学者分时期进行了研究，认为短期和长期结论不同，近代和现代结论不同。虽然学者们的研究结论存在种种争议，但他们的一个共识是金价和油价二者的变动之间存在着千丝万缕的关系。本章我们就用 Stata 14.0 分析研究一下原油和黄金的价格联动关系。

20.1 数据来源与研究思路

本章[1]所用的数据包括 WTI 自 2002 年 1 月 1 日至 2006 年 1 月 1 日，每月 1 日的原油价格数据共 49 组，LONDON GOLD FIX 自 2002 年 1 月至 2006 年 1 月的每月黄金价格均值数据共 49 组。其中原油价格数据来源于 http: / /www1forecasts1org / data / data /OILPRICE1htm，黄金价格数据来源于 http: / /www1forecasts1org / data / data /GOLD1htm。数据的 Excel 形式如表 20.1 所示。

表 20.1 案例 20 数据

month	lgoldf	wtioil	month	lgoldf	wtioil
1	281.65	19.67	26	405.33	34.74
2	295.5	20.74	27	406.67	36.76
3	294.05	24.42	28	403.02	36.69
4	302.68	26.27	29	383.4	40.28
5	314.49	27.02	30	391.99	38.02
6	310.25	25.52	31	398.09	40.69
7	313.29	26.94	32	400.48	44.94
8	310.25	28.38	33	405.25	45.95
9	319.16	29.67	34	423.34	53.13

[1] 改编自《石油与黄金产业价格联动关系研究》（由张莹、胥莉、陈宏民著），以及《财经问题研究》第7期（总第284期）。

（续表）

month	lgoldf	wtioil	month	lgoldf	wtioil
10	316.56	28.85	35	439.39	48.46
11	319.15	26.27	36	441.76	43.33
12	332.43	29.42	37	424.15	46.84
13	356.86	32.94	38	423.35	47.97
14	359.32	35.87	39	434.25	54.31
15	340.55	33.55	40	428.93	53.04
16	328.58	28.25	41	421.87	49.83
17	355.68	28.14	42	430.66	56.26
18	356.53	30.72	43	424.48	58.7
19	351	30.76	44	437.93	64.97
20	359.77	31.59	45	456.05	65.57
21	378.95	28.29	46	469.9	62.37
22	378.92	30.33	47	476.67	58.3
23	389.91	31.09	48	510.1	59.43
24	407.59	32.15	49	549.86	65.51
25	413.99	34.27			

本数据为时间序列数据，研究思路是首先对数据进行描述性分析，并绘制变量的时间序列趋势图，简明扼要地分析一下数据特征，进行相关性检验，探索变量之间的相关关系，然后对数据中两个时间序列采用多种方法进行单位根检验，综合分析其平稳性，再使用 EG-ADF 协整检验的方式对数据进行协整检验，综合分析其长期均衡关系，对两个变量进行格兰杰因果关系检验，探讨变量之间的格兰杰因果关系，最后建立相应的误差修正模型，并提出研究结论。

20.2　描述性分析

	下载资源:\video\chap20\⋯
	下载资源:\sample\chap20\案例20.dta

本案例的数据变量都是定距变量，通过进行定距变量的基本描述性统计可以得到数据的概要统计指标，包括平均值、最大值、最小值、标准差、百分位数、中位数、偏度系数和峰度系数等。我们通过获得这些指标，可以从整体上对拟分析的数据进行宏观把握，为后续进行更深入的数据分析做好必要准备。

20.2.1　Stata 分析过程

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有 3 个变量，分别为月份、原油价格和黄金价格。我们把月份变量设定为 month，把原油价格变量设定为 wtioil，把黄金价格变量设定为 lgoldf，变量类型及长度采取系统默认方式，然后录入相关数据。相关操作在第 1 章中已有详细讲述。录入完成后数据如图 20.1 所示。

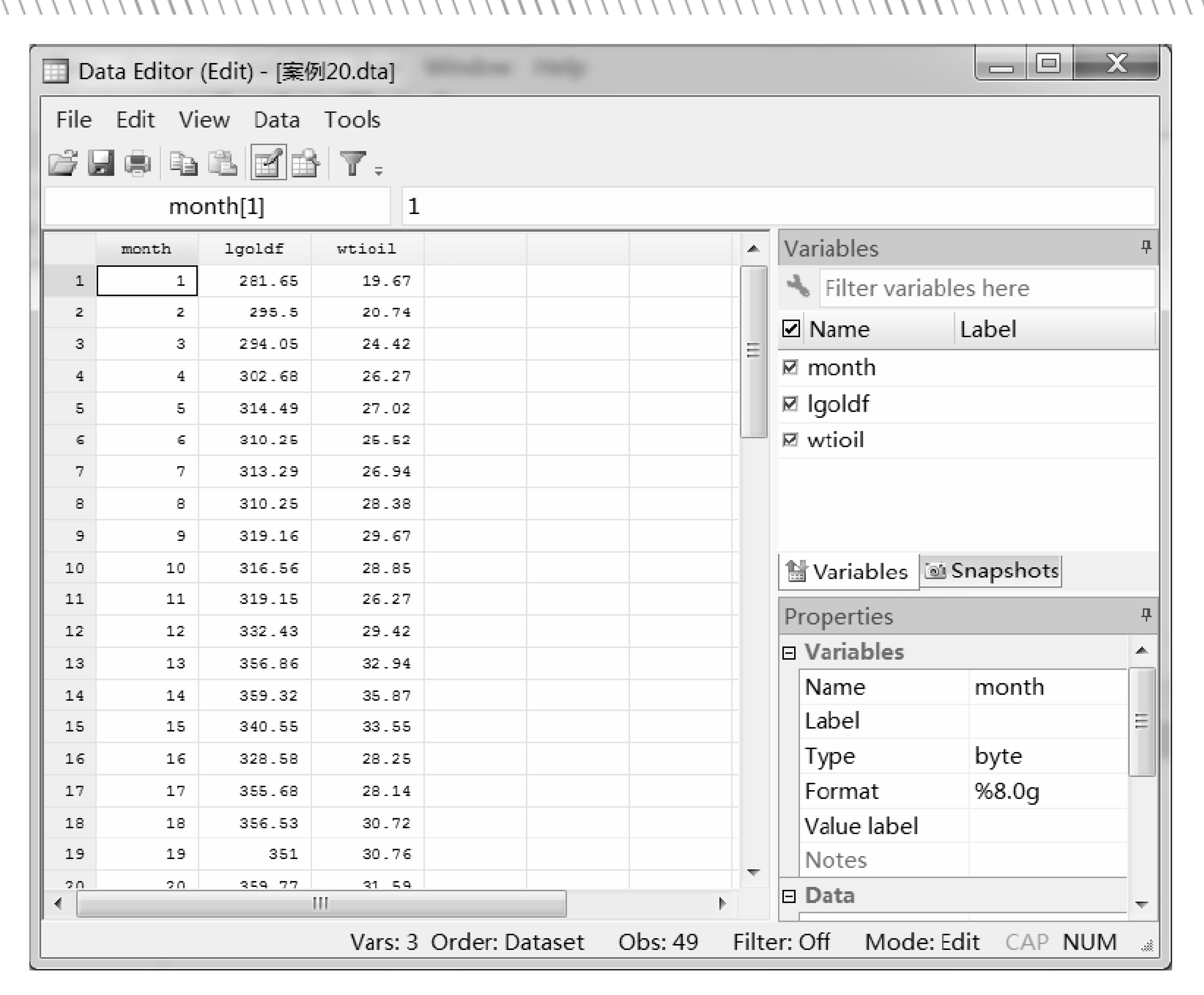


图 20.1 案例 20 数据

先做一下数据保存，然后开始展开分析，步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入命令：

```
summarize lgoldf wtioil,detail
```

03 设置完毕后，按键盘上的回车键，等待输出结果。

20.2.2 结果分析

在 Stata 14.0 主界面的结果窗口可以看到如图 20.2 所示的分析结果。在分析结果中，可以得到如下很多信息。

1．百分位数（Percentiles）

可以看出变量 lgoldf 的第 1 个四分位数(25%)是 332.43，第 2 个四分位数(50%)是 391.99，第 3 个四分位数（75%）是 424.15；变量 wtioil 的第 1 个四分位数（25%）是 28.85，第 2 个四分位数（50%）是 34.74，第 3 个四分位数（75%）是 48.46。

```
. summarize    lgoldf wtioil,detail

                          lgoldf
-------------------------------------------------------------
      Percentiles      Smallest
 1%       281.65         281.65
 5%        295.5         294.05
10%       310.25          295.5       Obs                  49
25%       332.43         302.68       Sum of Wgt.          49

50%       391.99                      Mean           385.1843
                        Largest       Std. Dev.      59.69529
75%       424.15          469.9
90%       456.05         476.67       Variance       3563.528
95%       476.67          510.1       Skewness       .3070091
99%       549.86         549.86       Kurtosis       2.804495

                          wtioil
-------------------------------------------------------------
      Percentiles      Smallest
 1%        19.67          19.67
 5%        24.42          20.74
10%        26.27          24.42       Obs                  49
25%        28.85          25.52       Sum of Wgt.          49

50%        34.74                      Mean           39.33082
                        Largest       Std. Dev.      13.02875
75%        48.46          62.37
90%        59.43          64.97       Variance       169.7483
95%        64.97          65.51       Skewness       .6038287
99%        65.57          65.57       Kurtosis       2.146637
```

图 20.2　分析结果图

2．4 个最小值（Smallest）

变量 lgoldf 最小的 4 个数据值分别是 281.65、294.05、295.5、302.68。
变量 wtioil 最小的 4 个数据值分别是 19.67、20.74、24.42、25.52。

3．4 个最大值（Largest）

变量 lgoldf 最大的 4 个数据值分别是 469.9、476.67、510.1、549.86。
变量 wtioil 最大的 4 个数据值分别是 62.37、64.97、65.51、65.57。

4．平均值（Mean）和标准差（Std. Dev）

变量 lgoldf 的平均值为 385.1843，标准差是 59.69529。
变量 wtioil 的平均值为 39.33082，标准差是 13.02875。

5．偏度（Skewness）和峰度（Kurtosis）

变量 lgoldf 的偏度为 0.3070091，为正偏度但不大。
变量 wtioil 的偏度为 0.6038287，为正偏度但不大。
变量 lgoldf 的峰度为 2.804495，有一个比正态分布更短的尾巴。
变量 wtioil 的峰度为 2.146637，有一个比正态分布更短的尾巴。

20.3　时间序列趋势图

通过绘制时间序列趋势图操作可以迅速地看出数据的变化特征，为后续更加精确地判断或者选择合适的模型做好必要准备。

20.3.1　Stata 分析过程

时间序列趋势图分析的步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入如下命令：

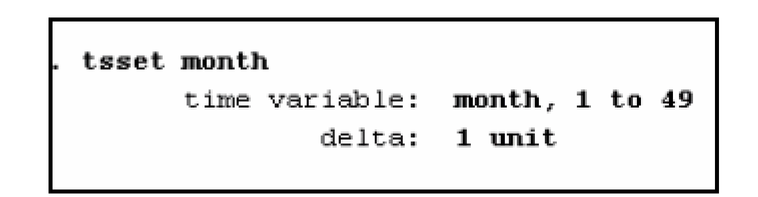
- tsset month
- twoway(line lgoldf month)
- twoway(line wtioil month)
- gen lnlgoldf=log(lgoldf)
- gen lnwtioil=log(wtioil)
- twoway(line lnlgoldf month)
- twoway(line lnwtioil month)
- twoway(line d.lnlgoldf month)
- twoway(line d.lnwtioil month)

03 设置完毕后，按键盘上的回车键，等待输出结果。

20.3.2 结果分析

在 Stata 14.0 主界面的结果窗口可以看到如图 20.3~图 20.11 所示的分析结果。

图 20.3 显示的是我们把月份作为日期变量对数据进行时间定义的结果。



```
. tsset month
        time variable:  month, 1 to 49
                delta:  1 unit
```

图 20.3 分析结果图 1

从上述分析结果中可以看到时间变量是月份（month），区间范围是 1~49，间距为 1。

图 20.4 显示的是变量黄金价格随时间的变动趋势。

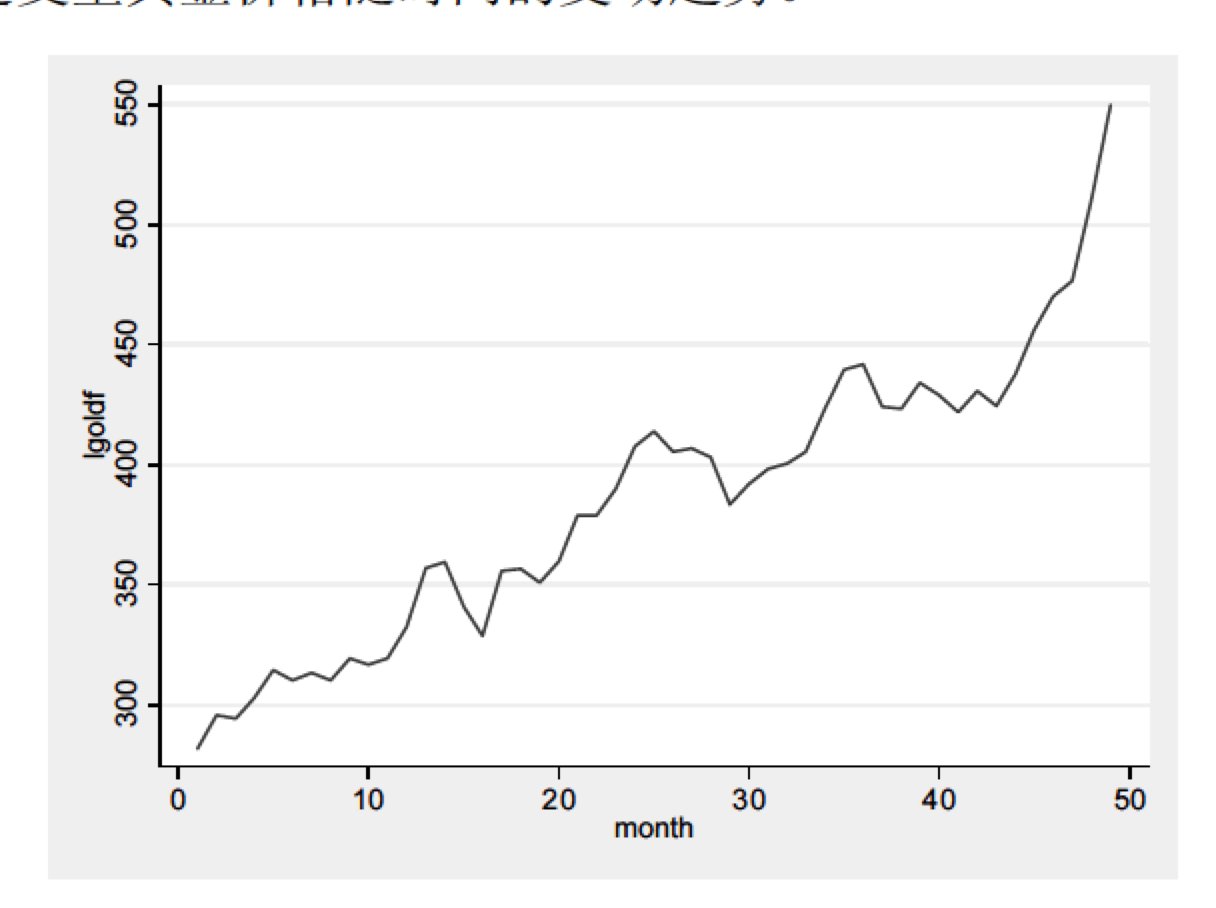


图 20.4 分析结果图 2

从上述分析结果中可以看到变量黄金价格具有明显、稳定的长期增长趋势。

图 20.5 显示的是变量原油价格随时间的变动趋势。

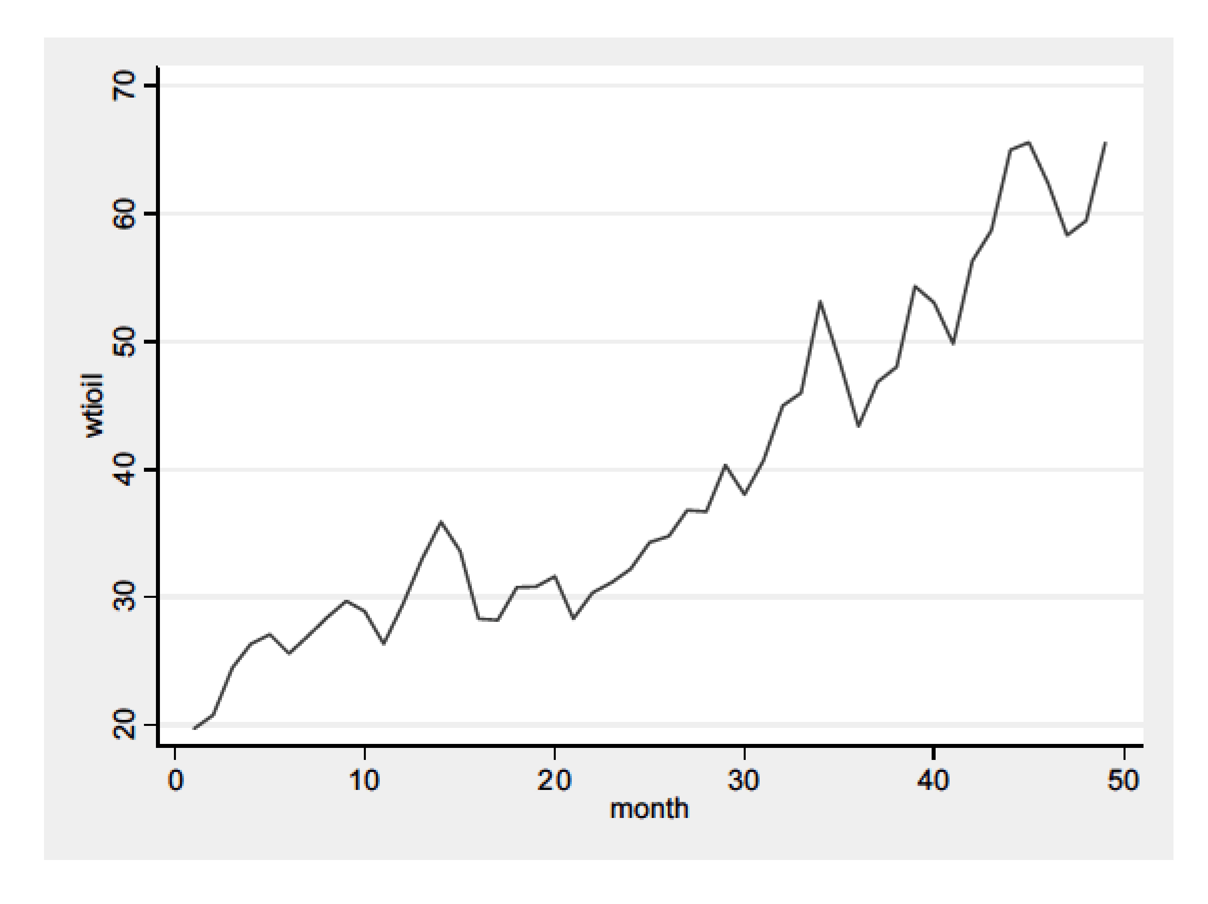


图 20.5　分析结果图 3

从上述分析结果中可以看到变量原油价格具有明显、稳定的长期增长趋势。

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 20.6 所示的 lnlgoldf 数据。lnlgoldf 数据是对数据 lgoldf 进行对数变换处理的结果，这步处理的意义是消除数据异方差的影响，使数据更适合深入分析，并且使数据更具实际意义。对数变换引出了弹性的概念，在没有进行对数变换之前，变量之间的联动关系表现在自变量的变动引起因变量变动的程度，在进行对数变换之后，变量的联动关系就表现为自变量变动的百分比引起因变量变动的百分比的程度。

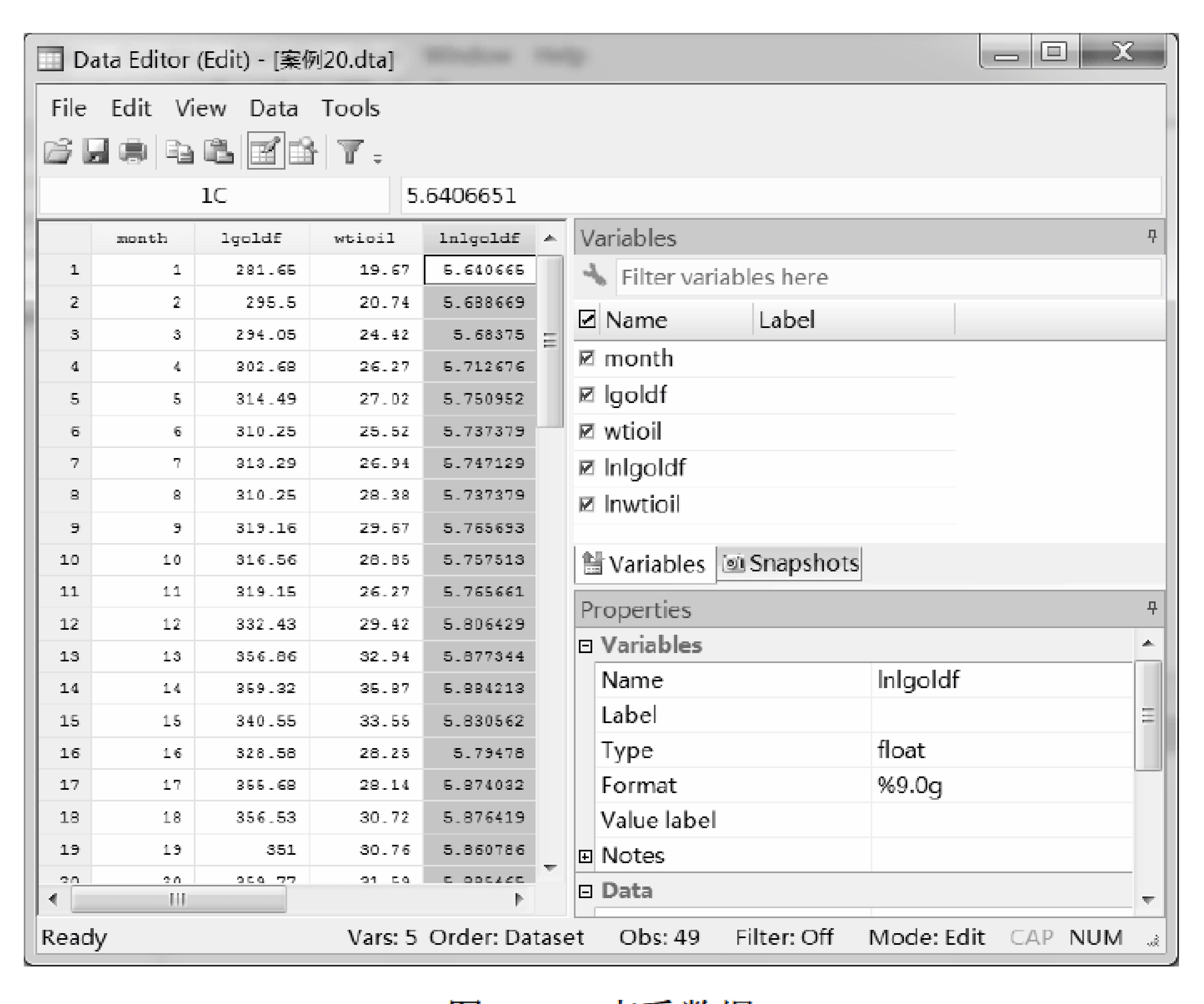


图 20.6　查看数据

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 20.7 所示的 lnwtioil 数据。lnwtioil 数据是对数据 wtioil 进行对数变换处理的结果。

图 20.8 显示的是变量黄金价格的对数值随时间的变动趋势。

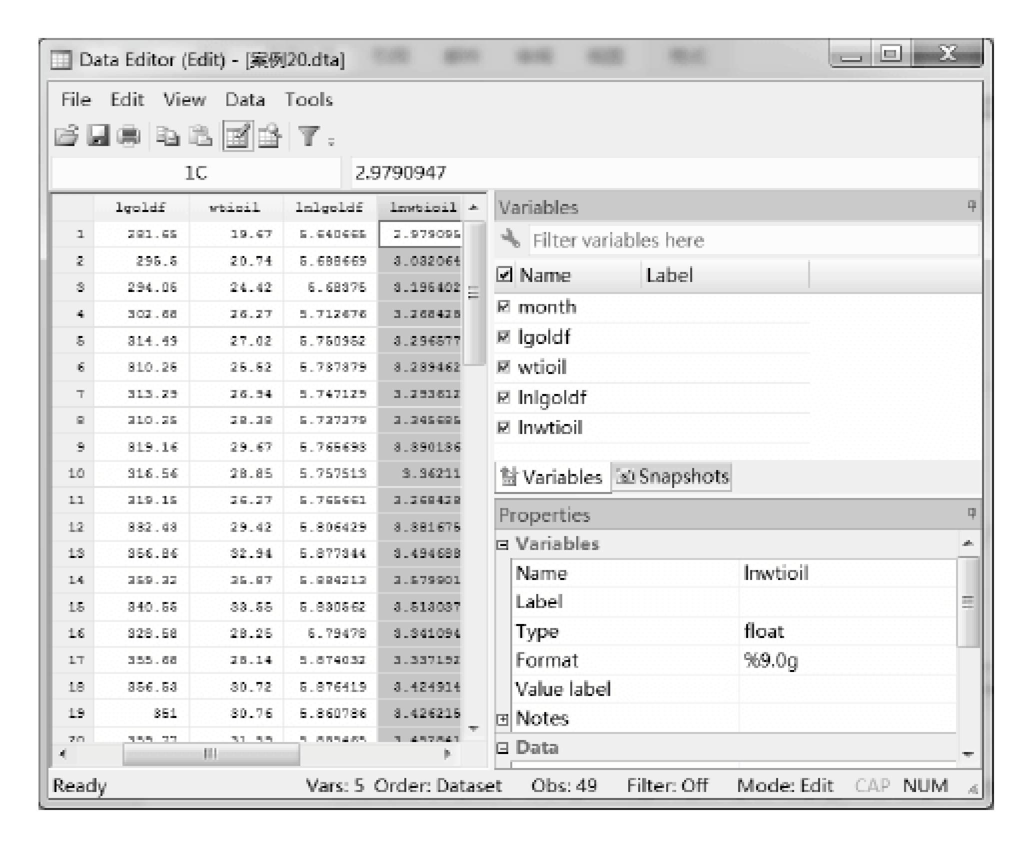


图 20.7　查看数据

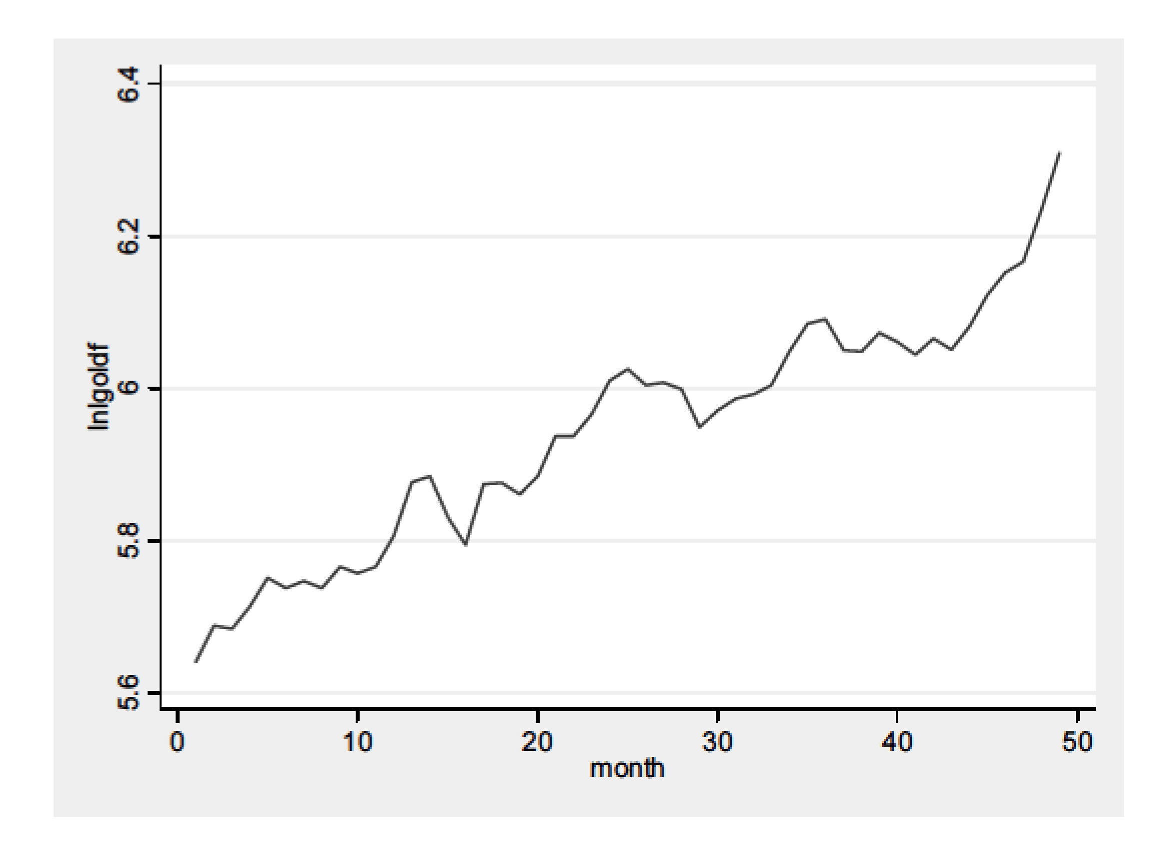


图 20.8　分析结果图 4

从上述分析结果中可以看到变量黄金价格的对数值具有明显、稳定的向上增长趋势。

图 20.9 显示的是变量原油价格的对数值随时间的变动趋势，从分析结果中可以看到原油价格的对数值具有明显、稳定的向上增长趋势。

图 20.10 显示的是变量黄金价格的对数值的一阶差分值随时间的变动趋势。

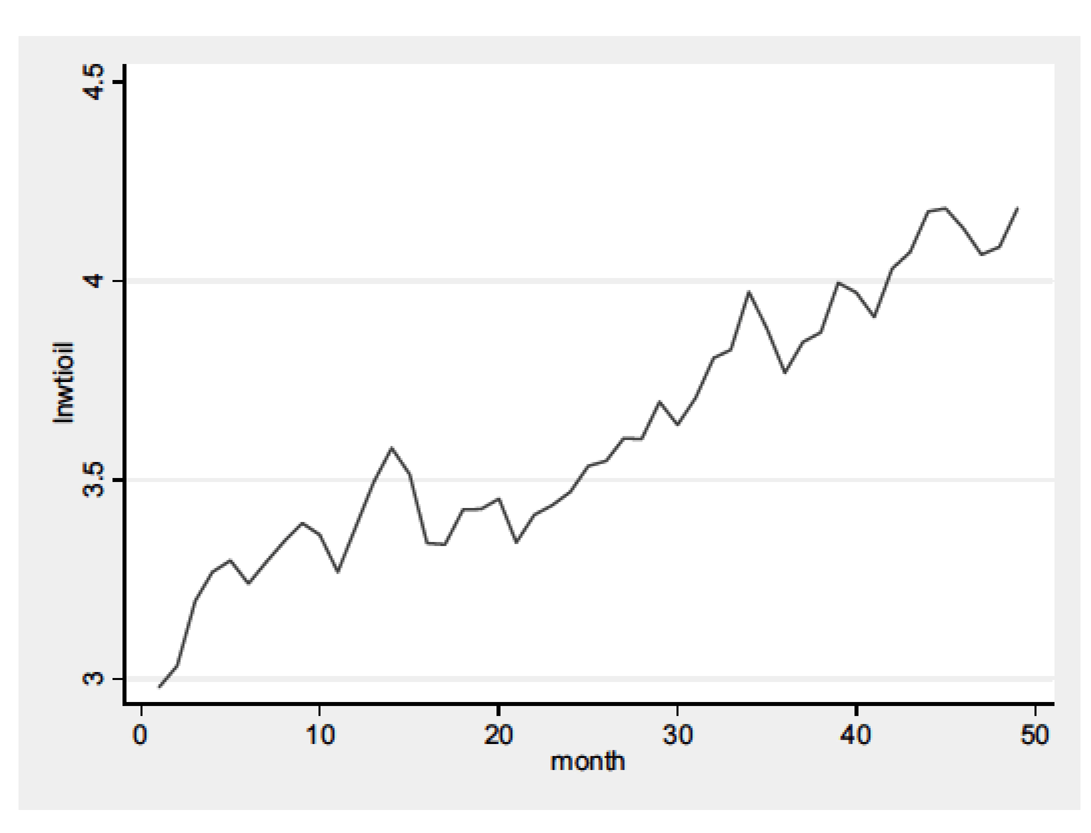


图 20.9　分析结果图 5

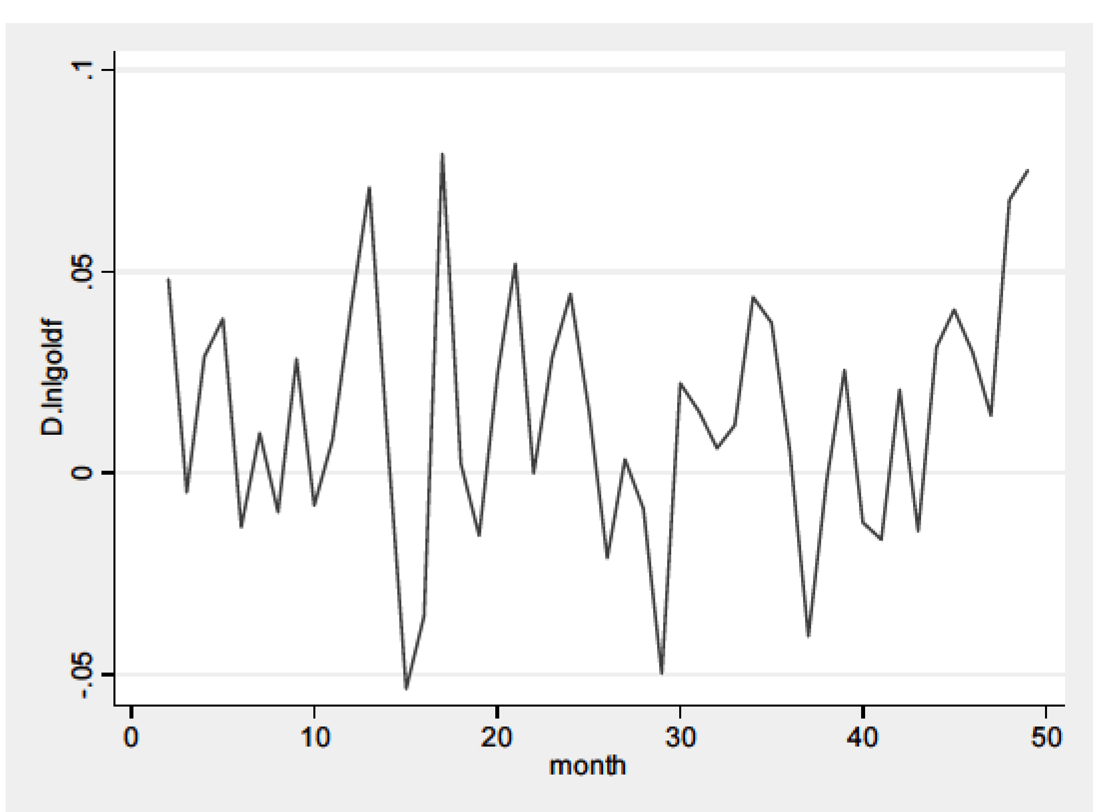


图 20.10　分析结果图 6

从上述分析结果中可以看到变量黄金价格的对数值的一阶差分值没有明显、稳定的长期变动趋势。

图 20.11 显示的是变量原油价格的对数值的一阶差分值随时间的变动趋势，从分析结果中可以看到变量原油价格的对数值的一阶差分值没有明显、稳定的长期变动趋势。

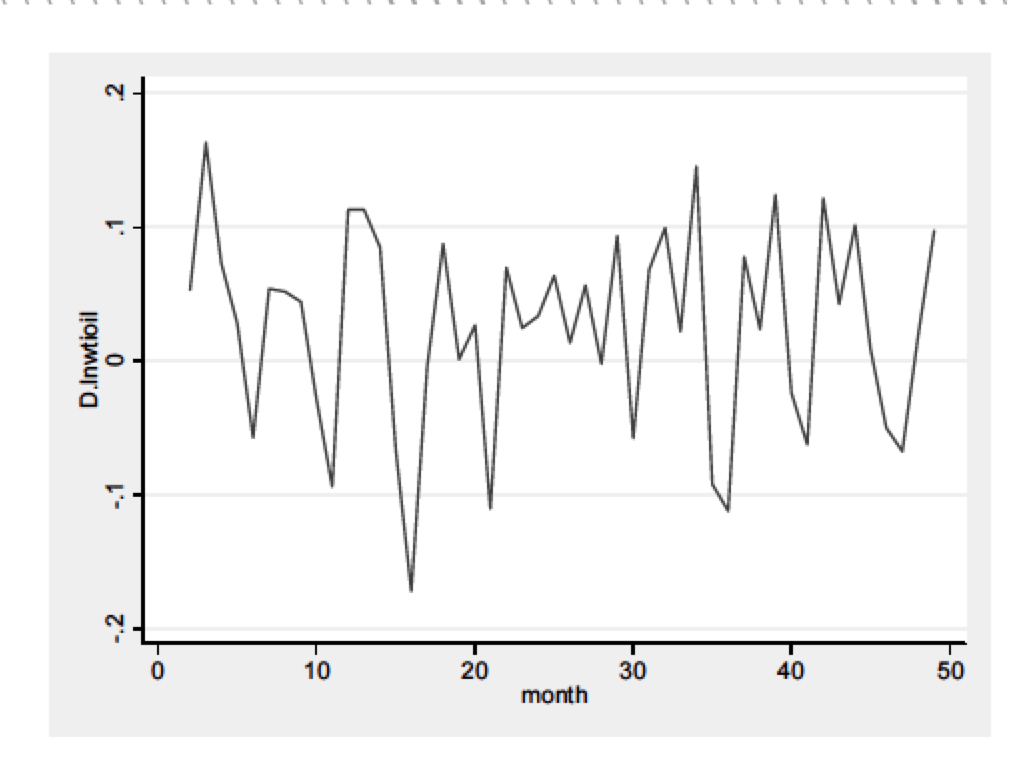


图 20.11　分析结果图 7

综上所述，我们通过绘制时间序列趋势图发现变量黄金价格的对数值的一阶差分值、原油价格的对数值的一阶差分值是没有时间趋势的，而变量黄金价格、原油价格、黄金价格的对数值、原油价格的对数值是有明显、稳定的向上增长趋势的。这些结论将会在后续的操作命令中被用到。

20.4　相关性分析

相关分析是不考虑变量之间的因果关系而只研究分析变量之间的相关关系的一种统计分析方法，通过该步操作我们可以判断出变量之间的相关性，从而考虑是否有必要进行后续分析或者增加替换新的变量等。

20.4.1　Stata 分析过程

相关性分析的步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入命令：

- correlate month lgoldf wtioil,covariance
- correlate month lnlgoldf lnwtioil,covariance
- correlate month lgoldf wtioil
- correlate month lnlgoldf lnwtioil
- pwcorr month lgoldf wtioil,sidak sig star(99)
- pwcorr month lnlgoldf lnwtioil,sidak sig star(99)

03 设置完毕后，按键盘上的回车键，等待输出结果。

20.4.2　结果分析

在 Stata 14.0 主界面的结果窗口可以看到如图 20.12~图 20.17 所示的分析结果。

图 20.12 展示的是变量黄金价格与原油价格的方差-协方差矩阵。

```
. correlate  month lgoldf wtioil,covariance
(obs=49)

             |    month   lgoldf   wtioil
-------------+---------------------------
       month |  204.167
      lgoldf |  815.722  3563.53
      wtioil |  175.321  695.264  169.748
```

图 20.12　分析结果图 1

从上述分析结果中可以看到月份的方差是 204.167，黄金价格的方差是 3563.53，石油价格的方差是 169.748，月份与黄金价格的协方差是 815.722，月份与石油价格的协方差是 175.321，黄金价格与石油价格之间的相关系数是 695.264。可以发现变量之间的方差差别是非常大的，我们对数据进行对数变换处理是非常有必要，也是非常有意义的。

图 20.13 展示的是变量黄金价格的对数值与原油价格的对数值的方差-协方差矩阵。

```
. correlate  month  lnlgoldf lnwtioil,covariance
(obs=49)

             |    month lnlgoldf lnwtioil
-------------+---------------------------
       month |  204.167
    lnlgoldf |  2.13546  .024054
    lnwtioil |  4.43194  .045727  .104746
```

图 20.13　分析结果图 2

从上述分析结果中可以看到月份的方差是 204.167，黄金价格对数值的方差是 0.024054，石油价格对数值的方差是 0.104746，月份与黄金价格对数值的协方差是 2.13546，月份与石油价格对数值的协方差是 4.43194，黄金价格对数值与石油价格对数值之间的相关系数是 0.045727。可以发现对变量进行对数变换处理后，变量的方差差距减少了很多，对数变换处理起到了应有的效果。

图 20.14 展示的是变量黄金价格与原油价格的相关系数矩阵。

```
. correlate month lgoldf wtioil
(obs=49)

             |    month   lgoldf   wtioil
-------------+---------------------------
       month |   1.0000
      lgoldf |   0.9563   1.0000
      wtioil |   0.9418   0.8939   1.0000
```

图 20.14　分析结果图 3

从上述分析结果中可以看到 3 个变量之间的相关系数非常高。其中月份与黄金价格之间的相关系数为 0.9563，月份与石油价格之间的相关系数为 0.9418。我们知道在本例中，变量月份的数据取值是从 1 开始到 49 的连续整数，黄金价格、石油价格与月份这一连续等距增长的数据有如此之高的正相关系数，说明这两个变量本身就是一种不断增长的趋势，这也在一定程度上验证了我们在时间序列趋势图阶段的分析结论。黄金价格与石油价格之间的相关系数为 0.8939，高的正相关系数在一定程度上说明这两个变量之间很可能存在着一定的联动关系，说明我们的后续分析是很有必要的。

图 20.15 展示的是变量黄金价格的对数值与原油价格的对数值的相关系数矩阵。

```
. correlate month lnlgoldf lnwtioil
(obs=49)

             |    month lnlgoldf lnwtioil
-------------+---------------------------
       month |   1.0000
    lnlgoldf |   0.9636   1.0000
    lnwtioil |   0.9584   0.9110   1.0000
```

图 20.15　分析结果图 4

从上述分析结果中可以看到经过对数变换处理以后，3 个变量之间的相关系数得到了进一步的提高。其中月份与黄金价格对数值之间的相关系数为 0.9636，月份与石油价格对数值之间的相关系数为 0.9584，黄金价格对数值与石油价格对数值之间的相关系数为 0.9110。

图 20.16 展示的是变量黄金价格与原油价格的相关系数矩阵的显著性检验，设定置信水平为 99%。

```
. pwcorr month  lgoldf wtioil,sidak sig star(99)

             |    month   lgoldf   wtioil
-------------+---------------------------
       month |   1.0000
             |
             |
      lgoldf |   0.9563*  1.0000
             |   0.0000
             |
      wtioil |   0.9418*  0.8939*  1.0000
             |   0.0000   0.0000
             |
```

图 20.16　分析结果图 5

从上述分析结果中可以看到 3 个变量之间的相关系数非常高，均通过了置信水平为 99%的相关性检验。

图 20.17 展示的是变量黄金价格的对数值与原油价格的对数值的相关系数矩阵的显著性检验，设定置信水平为 99%。

```
. pwcorr month  lnlgoldf lnwtioil,sidak sig star(99)

             |    month lnlgoldf lnwtioil
-------------+---------------------------
       month |   1.0000
             |
             |
    lnlgoldf |   0.9636*  1.0000
             |   0.0000
             |
    lnwtioil |   0.9584*  0.9110*  1.0000
             |   0.0000   0.0000
             |
```

图 20.17　分析结果图 6

从上述分析结果中可以看到 3 个变量经对数变换处理之后的相关系数依然非常高，均通过了置信水平为 99%的相关性检验。

20.5 单位根检验

对于时间序列数据而言，数据的平稳性对于模型的构建是非常重要的。如果时间序列数据是不平稳的，可能会导致自回归系数的估计值向左偏向于 0，使传统的 T 检验失效，也有可能会使得两个相互独立的变量出现假相关关系或者回归关系，造成模型结果的失真。单位根检验是判断数据是否平稳的重要方法。只有进行了该步操作，我们才能进行后续的深入分析。

20.5.1 Stata 分析过程

通过前面的分析可以发现经过对数变换处理之后的变量要优于原变量，所以我们在后续的分析中不再包含原变量，只针对对数变换之后的变量进行分析，并得出研究结论。本例我们采用 3 种单位根检验分析方法，分别是 PP 检验、ADF 检验以及 DF-GLS 检验。在前面我们通过绘制时间序列趋势图发现变量黄金价格的对数值的一阶差分值、原油价格的对数值的一阶差分值是没有时间趋势的，而变量黄金价格的对数值、原油价格的对数值是有明显、稳定的向上增长趋势的。这些结论将会在单位根检验的操作命令中被用到。

1. PP 检验

PP 检验的操作步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中分别输入如下命令并按键盘上的回车键进行确认：

- pperron lnlgoldf,trend
- pperron lnwtioil,trend
- pperron d.lnlgoldf,notrend
- pperron d.lnwtioil,notrend

03 设置完毕后，等待输出结果。

2. ADF 检验

ADF 检验的操作步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中分别输入如下命令并按键盘上的回车键进行确认：

- dfuller lnlgoldf,trend lags(1)
- dfuller lnwtioil,trend lags(2)
- dfuller d.lnlgoldf,notrend lags(1)
- dfuller d.lnwtioil,notrend lags(1)

03 设置完毕后，等待输出结果。

3．DF-GLS 检验

DF-GLS 检验的操作步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中分别输入如下命令并按键盘上的回车键进行确认：

- dfgls lnlgoldf
- dfgls　lnwtioil
- dfgls d.lnlgoldf,notrend
- dfgls d.lnwtioil,notrend

03 设置完毕后，等待输出结果。

20.5.2　结果分析

在 Stata 14.0 主界面的结果窗口可以看到如图 20.18~图 20.29 所示的分析结果。

1．PP 检验结果

PP 检验的结果如图 20.18~图 20.21 所示。其中，图 20.18 展示的是黄金价格的对数值这一变量的 PP 检验结果。

```
. pperron  lnlgoldf,trend

Phillips-Perron test for unit root                 Number of obs    =        48
                                                   Newey-West lags =         3

                         ---------- Interpolated Dickey-Fuller ---------
                  Test       1% Critical      5% Critical      10% Critical
               Statistic        Value            Value             Value
------------------------------------------------------------------------------
 Z(rho)         -13.964         -25.444          -19.648           -16.704
 Z(t)            -2.343          -4.168           -3.508            -3.185
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.4103
```

图 20.18　分析结果图 1

PP 检验的原假设是数据有单位根。从上面的结果中可以看出 P 值(MacKinnon approximate p-value for Z(t)）为 0.4103，接受了有单位根的原假设，这一点也可以通过观察 Z(t)值和 Z(rho)值得到。实际 Z(t)值为-2.343，在 1%的置信水平（-4.168）、5%的置信水平（-3.508）、10%的置信水平上(-3.185)都无法拒绝原假设。实际 Z(rho)值为-13.964，在 1%的置信水平(-25.444)、5%的置信水平（-19.648）、10%的置信水平上（-16.704）都无法拒绝原假设，所以黄金价格的对数值这一变量数据是存在单位根的，需要对其做一阶差分后再继续进行检验。

图 20.19 展示的是原油价格的对数值这一变量的 PP 检验结果。

```
. pperron   lnwtioil,trend

Phillips-Perron test for unit root                 Number of obs    =        48
                                                   Newey-West lags =         3

                         ---------- Interpolated Dickey-Fuller ---------
                  Test       1% Critical      5% Critical      10% Critical
               Statistic        Value            Value             Value
------------------------------------------------------------------------------
 Z(rho)         -16.484         -25.444          -19.648           -16.704
 Z(t)            -3.149          -4.168           -3.508            -3.185
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t| = 0.0950
```

图 20.19　分析结果图 2

PP 检验的原假设是数据有单位根。从上面的结果中可以看出 P 值（MacKinnon approximate p-value for Z(t)）为 0.0950，接受了有单位根的原假设，这一点也可以通过观察 Z(t)值和 Z(rho)值得到。实际 Z(t)值为-3.149，在 1%的置信水平（-4.168）、5%的置信水平（-3.508）、10%的置信水平上（-3.185）都无法拒绝原假设。实际 Z(rho)值为-16.484，在 1%的置信水平（-25.444）、5%的置信水平（-19.648）、10%的置信水平上（-16.704）都无法拒绝原假设，所以原油价格的对数值这一变量数据是存在单位根的，需要对其做一阶差分后再继续进行检验。

图 20.20 展示的是黄金价格的对数值的一阶差分值这一变量的 PP 检验结果。

```
. pperron  d.lnlgoldf,notrend

Phillips-Perron test for unit root                  Number of obs    =        47
                                                    Newey-West lags  =         3

                               ---------- Interpolated Dickey-Fuller ---------
                  Test          1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(rho)          -34.849          -18.696           -13.204           -10.640
 Z(t)             -5.440           -3.600            -2.938            -2.604
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0000
```

图 20.20　分析结果图 3

PP 检验的原假设是数据有单位根。从上面的结果中可以看出 P 值（MacKinnon approximate p-value for Z(t)）为 0.0000，拒绝了有单位根的原假设，这一点也可以通过观察 Z(t)值和 Z(rho)值得到。实际 Z(t)值为-5.440，在 1%的置信水平（-3.600）、5%的置信水平（-2.938）、10%的置信水平上（-2.604）都拒绝了原假设。实际 Z(rho)值为-34.849，在 1%的置信水平（-18.696）、5%的置信水平（-13.204）、10%的置信水平上（-10.640）都拒绝了原假设，所以黄金价格的对数值的一阶差分值这一变量数据是不存在单位根的。

图 20.21 展示的是原油价格的对数值的一阶差分值这一变量的 PP 检验结果。

```
. pperron  d.lnwtioil,notrend

Phillips-Perron test for unit root                  Number of obs    =        47
                                                    Newey-West lags  =         3

                               ---------- Interpolated Dickey-Fuller ---------
                  Test          1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(rho)          -35.177          -18.696           -13.204           -10.640
 Z(t)             -6.434           -3.600            -2.938            -2.604
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0000
```

图 20.21　分析结果图 4

PP 检验的原假设是数据有单位根。从上面的结果中可以看出 P 值（MacKinnon approximate p-value for Z(t)）为 0.0000，拒绝了有单位根的原假设，这一点也可以通过观察 Z(t)值和 Z(rho)值得到。实际 Z(t)值为-6.434，在 1%的置信水平（-3.600）、5%的置信水平（-2.938）、10%的置信水平上（-2.604）都拒绝了原假设。实际 Z(rho)值为-35.177，在 1%的置信水平（-18.696）、5%的置信水平（-13.204）、10%的置信水平上（-10.640）都拒绝了原假设，所以原油价格的对数值的一阶差分值这一变量数据是不存在单位根的。

2．ADF 检验结果

ADF 检验的结果如图 20.22~图 20.25 所示。其中，图 20.22 展示的是黄金价格的对数值这

一变量的 ADF 检验结果。

```
. dfuller lnlgoldf,trend lags(1)

Augmented Dickey-Fuller test for unit root         Number of obs   =        47

                               ---------- Interpolated Dickey-Fuller ----------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -2.548            -4.178            -3.512            -3.187
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.3043
```

图 20.22　分析结果图 5

ADF 检验的原假设是数据有单位根。从上面的结果中可以看出 P 值（MacKinnon approximate p-value for Z(t)）为 0.3043，接受了有单位根的原假设，这一点也可以通过观察 Z(t)值和 Z(rho)值得到。实际 Z(t)值为-2.548，在 1%的置信水平（-4.178）、5%的置信水平（-3.512）、10%的置信水平上（-3.187）都无法拒绝原假设，所以黄金价格的对数值这一变量数据是存在单位根的，需要对其做一阶差分后再继续进行检验。

图 20.23 展示的是原油价格的对数值这一变量的 ADF 检验结果。

```
. dfuller lnwtioil,trend lags(2)

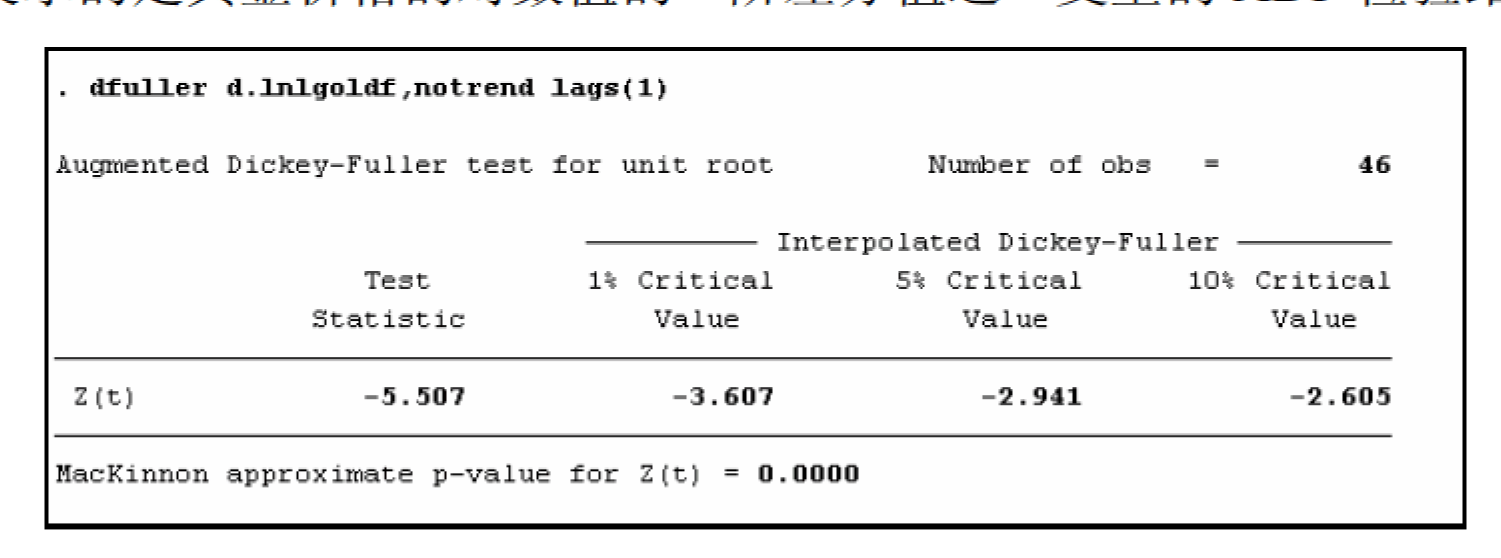
Augmented Dickey-Fuller test for unit root         Number of obs   =        46

                               ---------- Interpolated Dickey-Fuller ----------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -2.674            -4.187            -3.516            -3.190
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.2469
```

图 20.23　分析结果图 6

ADF 检验的原假设是数据有单位根。从上面的结果中可以看出 P 值（MacKinnon approximate p-value for Z(t)）为 0.2469，接受了有单位根的原假设，这一点也可以通过观察 Z(t)值和 Z(rho)值得到。实际 Z(t)值为-2.674，在 1%的置信水平（-4.187）、5%的置信水平（-3.516）、10%的置信水平上（-3.190）都无法拒绝原假设，所以原油价格的对数值这一变量数据是存在单位根的，需要对其做一阶差分后再继续进行检验。

图 20.24 展示的是黄金价格的对数值的一阶差分值这一变量的 ADF 检验结果。



```
. dfuller d.lnlgoldf,notrend lags(1)

Augmented Dickey-Fuller test for unit root         Number of obs   =        46

                               ---------- Interpolated Dickey-Fuller ----------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -5.507            -3.607            -2.941            -2.605
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0000
```

图 20.24　分析结果图 7

ADF 检验的原假设是数据有单位根。从上面的结果中可以看出 P 值（MacKinnon approximate p-value for Z(t)）为 0.0000，拒绝了有单位根的原假设，这一点也可以通过观察 Z(t)

值和 Z(rho)值得到。实际 Z(t)值为-5.507，在 1%的置信水平（-3.607）、5%的置信水平（-2.941）、10%的置信水平上（-2.605）都拒绝了原假设，所以黄金价格的对数值的一阶差分值这一变量数据是不存在单位根的。

图 20.25 展示的是原油价格的对数值的一阶差分值这一变量的 ADF 检验结果。

```
. dfuller d.lnwtioil,notrend lags(1)

Augmented Dickey-Fuller test for unit root         Number of obs   =        46

                               ---------- Interpolated Dickey-Fuller ----------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -6.154            -3.607            -2.941            -2.605
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.0000
```

图 20.25　分析结果图 8

ADF 检验的原假设是数据有单位根。从上面的结果中可以看出 P 值（MacKinnon approximate p-value for Z(t)）为 0.0000，拒绝了有单位根的原假设，这一点也可以通过观察 Z(t)值和 Z(rho)值得到。实际 Z(t)值为-6.154，在 1%的置信水平（-3.607）、5%的置信水平（-2.941）、10%的置信水平上（-2.605）都拒绝了原假设，所以原油价格的对数值的一阶差分值这一变量数据是不存在单位根的。

3. DF-GLS 检验结果

DF-GLS 检验的结果如图 20.26~图 20.29 所示。其中，图 20.26 展示的是黄金价格的对数值这一变量的 DF-GLS 检验结果。

```
. dfgls lnlgoldf

DF-GLS for lnlgoldf                                      Number of obs =    38
Maxlag = 10 chosen by Schwert criterion

               DF-GLS tau      1% Critical       5% Critical      10% Critical
  [lags]     Test Statistic        Value             Value             Value
------------------------------------------------------------------------------
    10           -1.459           -3.770            -2.673            -2.366
     9           -1.229           -3.770            -2.723            -2.425
     8           -1.434           -3.770            -2.783            -2.490
     7           -1.563           -3.770            -2.850            -2.559
     6           -2.119           -3.770            -2.921            -2.630
     5           -2.005           -3.770            -2.994            -2.701
     4           -2.678           -3.770            -3.066            -2.769
     3           -2.271           -3.770            -3.133            -2.833
     2           -1.681           -3.770            -3.195            -2.889
     1           -2.470           -3.770            -3.247            -2.937

Opt Lag (Ng-Perron seq t) =  3 with RMSE  .0269081
Min SC   =  -6.84775 at lag  3 with RMSE  .0269081
Min MAIC = -6.791573 at lag  2 with RMSE  .0285596
```

图 20.26　分析结果图 9

DF-GLS 检验的原假设是数据有单位根。从上面的结果中可以看出根据信息准则确定的最优滞后阶数为 3 阶（Opt Lag (Ng-Perron seq t) = 3 with RMSE.0269081），在该阶数下 DF-GLS 统计量的值是-2.271，在 1%的置信水平（-3.770）、5%的置信水平（-3.133）、10%的置信水平上（-2.833）都无法拒绝原假设，所以黄金价格的对数值这一变量数据是存在单位根的，需要对其做一阶差分后再继续进行检验。

图 20.27 展示的是原油价格的对数值这一变量的 ADF 检验结果。

```
. dfgls  lnwtioil

DF-GLS for lnwtioil                                     Number of obs =    38
Maxlag = 10 chosen by Schwert criterion

               DF-GLS tau      1% Critical       5% Critical      10% Critical
  [lags]     Test Statistic       Value             Value             Value
------------------------------------------------------------------------------
    10          -2.047           -3.770            -2.673            -2.366
     9          -2.110           -3.770            -2.723            -2.425
     8          -1.691           -3.770            -2.783            -2.490
     7          -1.693           -3.770            -2.850            -2.559
     6          -1.842           -3.770            -2.921            -2.630
     5          -1.968           -3.770            -2.994            -2.701
     4          -1.386           -3.770            -3.066            -2.769
     3          -1.854           -3.770            -3.133            -2.833
     2          -2.522           -3.770            -3.195            -2.889
     1          -3.068           -3.770            -3.247            -2.937

Opt Lag (Ng-Perron seq t) =  5 with RMSE  .0539539
Min SC   = -5.264894 at lag  5 with RMSE  .0539539
Min MAIC = -5.148633 at lag  4 with RMSE  .0614735
```

图 20.27　分析结果图 10

DF-GLS 检验的原假设是数据有单位根。从上面的结果中可以看出根据信息准则确定的最优滞后阶数为 5 阶（Opt Lag (Ng-Perron seq t) = 5 with RMSE 0.0539539），在该阶数下 DF-GLS 统计量的值是-1.968，在 1%的置信水平（-3.770）、5%的置信水平（-2.994）、10%的置信水平上（-2.701）都无法拒绝原假设，所以原油价格的对数值这一变量数据是存在单位根的，需要对其做一阶差分后再继续进行检验。

图 20.28 展示的是黄金价格的对数值的一阶差分值这一变量的 DF-GLS 检验结果。

```
. dfgls d.lnlgoldf,notrend

DF-GLS for D.lnlgoldf                                   Number of obs =    38
Maxlag = 9 chosen by Schwert criterion

               DF-GLS mu       1% Critical       5% Critical      10% Critical
  [lags]     Test Statistic       Value             Value             Value
------------------------------------------------------------------------------
     9          -0.886           -2.623            -2.087            -1.778
     8          -1.143           -2.623            -2.101            -1.798
     7          -1.239           -2.623            -2.124            -1.824
     6          -1.475           -2.623            -2.152            -1.854
     5          -1.408           -2.623            -2.185            -1.888
     4          -1.699           -2.623            -2.221            -1.923
     3          -1.526           -2.623            -2.256            -1.958
     2          -1.961           -2.623            -2.290            -1.990
     1          -3.705           -2.623            -2.321            -2.018

Opt Lag (Ng-Perron seq t) =  2 with RMSE  .0302886
Min SC   = -6.706793 at lag  2 with RMSE  .0302886
Min MAIC = -6.609422 at lag  9 with RMSE   .026814
```

图 20.28　分析结果图 11

DF-GLS 检验的原假设是数据有单位根。从上面的结果中可以看出根据信息准则确定的最优滞后阶数为 2 阶（Opt Lag (Ng-Perron seq t) = 2 with RMSE 0.0302886），在该阶数下 DF-GLS 统计量的值是-1.961，在 1%的置信水平（-2.623）、5%的置信水平（-2.290）、10%的置信水平上（-1.990）都无法拒绝原假设，所以黄金价格的对数值的一阶差分值这一变量数据是存在单位根的。这一点显然与我们前面的检验结果不一致，但是这也是正常情况，事实上我们选择多种检验方法对数据进行单位根检验的初衷就是综合各种检验方法的检验结果做出恰当的判断。

图 20.29 展示的是原油价格的对数值的一阶差分值这一变量的 DF-GLS 检验结果。

```
. dfgls d.lnwtioil,notrend

DF-GLS for D.lnwtioil                                    Number of obs =     38
Maxlag = 9 chosen by Schwert criterion

               DF-GLS mu      1% Critical       5% Critical      10% Critical
  [lags]    Test Statistic       Value             Value             Value
------------------------------------------------------------------------------
    9           -0.987           -2.623            -2.087            -1.778
    8           -1.012           -2.623            -2.101            -1.798
    7           -1.434           -2.623            -2.124            -1.824
    6           -1.655           -2.623            -2.152            -1.854
    5           -1.747           -2.623            -2.185            -1.888
    4           -1.842           -2.623            -2.221            -1.923
    3           -4.508           -2.623            -2.256            -1.958
    2           -5.016           -2.623            -2.290            -1.990
    1           -5.217           -2.623            -2.321            -2.018

Opt Lag (Ng-Perron seq t) =  4 with RMSE  .0599442
Min SC    = -5.150052 at lag  4 with RMSE  .0599442
Min MAIC  = -4.494275 at lag  8 with RMSE  .0573994
```

图 20.29　分析结果图 12

DF-GLS 检验的原假设是数据有单位根。从上面的结果中可以看出根据信息准则确定的最优滞后阶数为 4 阶（Opt Lag (Ng-Perron seq t) = 4 with RMSE 0.0599442），在该阶数下 DF-GLS 统计量的值是-1.842，在 1%的置信水平（-2.623）、5%的置信水平（-2.221）、10%的置信水平上（-1.923）都无法拒绝原假设，所以原油价格的对数值的一阶差分值这一变量数据是存在单位根的。这一点显然与我们前面的检验结果不一致，但是这也是正常情况。

根据以上的分析，综合考虑 3 种检验方法的检验结果，可以比较有把握地得出以下结论，即认为变量黄金价格的对数值、原油价格的对数值是存在单位根的，黄金价格的对数值的一阶差分值、原油价格的对数值的一阶差分值是不存在单位根的，变量黄金价格的对数值、原油价格的对数值是一阶单整的。在该结论的基础上，将进入下一步的协整检验分析过程。

20.6　协整检验

在时间序列数据不平稳的情况下，构建出合理模型的重要方法就是进行协整检验并构建合理模型的处理方式。协整的思想就是把存在一阶单整的变量放在一起进行分析，通过这些变量做线性组合，从而消除它们的随机趋势，得到其长期联动趋势。

20.6.1　Stata 分析过程

本例采用 EG-ADF 协整检验分析方法进行分析。在前面的小节中，我们通过绘制时间序列趋势图发现变量黄金价格的对数值的一阶差分值、原油价格的对数值的一阶差分值是没有时间趋势的，而变量黄金价格的对数值、原油价格的对数值是有明显、稳定的向上增长趋势的。通过 PP 检验、ADF 检验以及 DF-GLS 检验等单位根检验发现变量黄金价格的对数值、原油价格的对数值是存在单位根的，黄金价格的对数值的一阶差分值、原油价格的对数值的一阶差分值是不存在单位根的，变量黄金价格的对数值、原油价格的对数值是一阶单整的。这些结论将会在协整检验的操作命令中被用到。

本例 EG-ADF 检验的操作步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中分别输入如下命令并按键盘上的回车键进行确认：

- reg lnlgoldf lnwtioil
- predict e,resid
- twoway(line e month)
- dfuller e,notrend nocon lags(1) regress
- reg e lnlgoldf lnwtioil

03 设置完毕后，等待输出结果。

20.6.2 结果分析

在 Stata 14.0 主界面的结果窗口可以看到如图 20.30~图 20.33 所示的分析结果。

本例 EG-ADF 检验过程是这样的：首先把黄金价格的对数值作为因变量，把原油价格的对数值作为自变量，用普通最小二乘估计法进行估计得到残差序列，然后对残差序列进行 ADF 检验，观测其是否为平稳序列，如果残差序列是平稳的，那么变量之间的长期协整关系就存在，如果残差序列是不平稳的，那么变量之间的长期协整关系就不存在。

图 20.30 展示的是把黄金价格的对数值作为因变量，把原油价格的对数值作为自变量，用普通最小二乘估计法进行估计的结果。

```
. reg  lnlgoldf lnwtioil

      Source |       SS           df       MS      Number of obs   =        49
-------------+----------------------------------   F(1, 47)        =    229.28
       Model |  .958171095         1  .958171095   Prob > F        =    0.0000
    Residual |  .196417615        47  .004179098   R-squared       =    0.8299
-------------+----------------------------------   Adj R-squared   =    0.8263
       Total |  1.15458871        48  .024053931   Root MSE        =    .06465

------------------------------------------------------------------------------
    lnlgoldf |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lnwtioil |   .4365488   .0288305    15.14   0.000     .3785492    .4945483
       _cons |   4.361619   .1047776    41.63   0.000     4.150834    4.572404
------------------------------------------------------------------------------
```

图 20.30 分析结果图 1

从上述分析结果中可以得到很多信息。可以看出共有 49 个样本参与了分析，模型的 F 值 (1, 47) =229.28，P 值（Prob > F）= 0.0000，说明模型整体上是非常显著的。模型的可决系数（R-squared）为 0.8299，模型修正的可决系数（Adj R-squared）为 0.8263，说明模型的解释能力还是差强人意的。

模型的回归方程是：

lnlgoldf =0.4365488* lnwtioil +4.361619

变量 lnwtioil 的系数标准误是 0.0288305，t 值为 15.14，P 值为 0.000，系数是非常显著的，95%的置信区间为[0.3785492 , 0.4945483]。常数项的系数标准误是 0.1047776，t 值为 41.63，P

值为 0.000，系数也是非常显著的，95%的置信区间为[4.150834 , 4.572404]。

从上面的分析可以看出简单回归的模型在一定程度上是可以接受的，但也存在提升改进的空间。本模型得到的基本结论是黄金价格和石油价格是一种正向联动关系，石油价格的升高会带来黄金价格的升高。

图 20.31 展示的是对模型残差的预测结果。选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 20.31 所示的数据。

	month	lgoldf	wtioil	lnlgoldf	lnwtioil	e
1	1	281.65	19.67	5.640665	2.979095	-.021474
2	2	295.5	20.74	5.688669	3.032064	.0034059
3	3	294.05	24.42	5.68375	3.195402	-.0728182
4	4	302.68	26.27	5.712676	3.268428	-.0757709
5	5	314.49	27.02	5.750952	3.296577	-.0497835
6	6	310.25	25.52	5.737379	3.239462	-.0384236
7	7	313.29	26.94	5.747129	3.293612	-.0523119
8	8	310.25	28.38	5.737379	3.345685	-.0847948
9	9	319.16	29.67	5.765693	3.390136	-.0758861
10	10	316.56	28.85	5.757513	3.36211	-.0718313
11	11	319.15	26.27	5.765661	3.268428	-.0227857
12	12	332.43	29.42	5.806429	3.381675	-.0314555
13	13	356.86	32.94	5.877344	3.494688	-.0098769
14	14	359.32	35.87	5.884213	3.579901	-.040207
15	15	340.55	33.55	5.830562	3.513037	-.0646687
16	16	328.58	28.25	5.79478	3.341094	-.025389
17	17	355.68	28.14	5.874032	3.337192	.0555655
18	18	356.53	30.72	5.876419	3.424914	.0196577
19	19	351	30.76	5.860786	3.426215	.0034575
20	20	359.77	31.59	5.885465	3.452841	.0165129
21	21	378.95	28.29	5.937404	3.342508	.1166173
22	22	378.92	30.33	5.937325	3.412137	.0861417
23	23	389.91	31.09	5.965916	3.436886	.1039287
24	24	407.59	32.15	6.010262	3.470412	.1336383
25	25	413.99	34.27	6.025842	3.53427	.1213414
26	26	405.33	34.74	6.004702	3.547892	.0942549
27	27	406.67	36.76	6.008002	3.60441	.0728825
28	28	403.02	36.69	5.998986	3.602504	.0646985
29	29	383.4	40.28	5.949079	3.695855	-.0259609
30	30	391.99	38.02	5.971236	3.638112	.0214038

图 20.31　分析结果图 2

图 20.32 展示的是残差序列的时间走势，可以发现残差序列是没有固定时间趋势的。

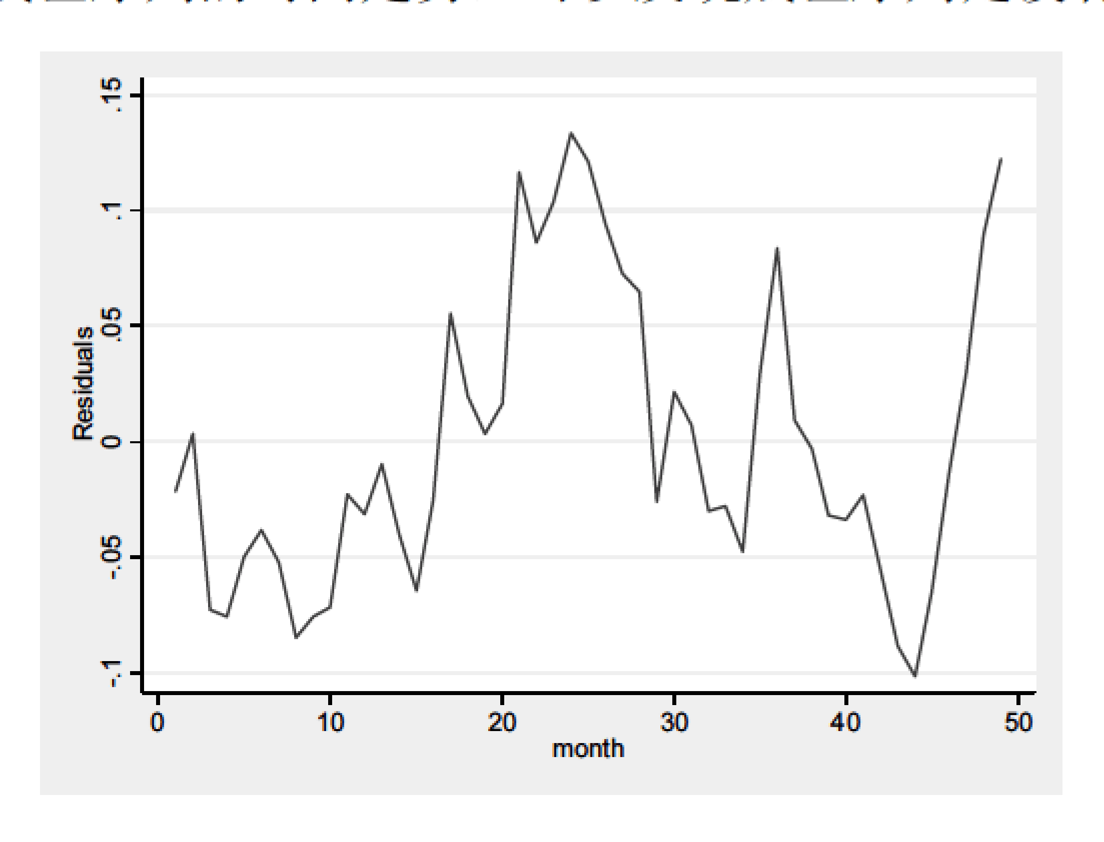


图 20.32　分析结果图 3

图 20.33 展示的是残差序列 ADF 检验结果。

```
. dfuller e,notrend nocon lags(1) regress

Augmented Dickey-Fuller test for unit root         Number of obs   =        47

                               ---------- Interpolated Dickey-Fuller ----------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -2.052            -2.625            -1.950            -1.609

------------------------------------------------------------------------------
         D.e |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           e |
         L1. |   -.203746   .0992795    -2.05   0.046    -.4037052   -.0037868
         LD. |   .1433098   .1529618     0.94   0.354     -.164771    .4513906
------------------------------------------------------------------------------
```

图 20.33　分析结果图 4

ADF 检验的原假设是数据有单位根。从上面的结果中可以看出实际 Z(t)值为-2.052，介于1%的置信水平（-2.625）和 5%的置信水平（-1.950）之间，所以在 5%的显著性水平上应该拒绝存在单位根的原假设，残差序列是不存在单位根的，或者说残差序列是平稳的。

综上所述，黄金价格的对数值、原油价格的对数值两个变量间存在一定的协整关系。根据上面的分析结果可以构建出相应的模型来描述这种协整关系。这一点我们在后续章节中将有详细说明。

20.7　格兰杰因果关系检验

协整关系表示的仅仅是变量之间的某种长期联动关系，与因果关系是毫无关联的，例如本例中虽然黄金价格与原油价格之间存在协整关系，但是究竟是黄金价格影响了原油价格，还是原油价格影响了黄金价格，亦或是它们相互影响？如果要探究变量之间的因果关系，就需要用到格兰杰因果关系检验。

20.7.1　Stata 分析过程

在前面几节中，通过单位根检验发现黄金价格的对数值、原油价格的对数值两个变量是一阶单整的，所以我们在进行格兰杰因果关系检验时选择的变量是：黄金价格的对数值、原油价格的对数值。

格兰杰因果关系检验的操作步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中分别输入如下命令并按键盘上的回车键进行确认：

- reg　lnlgoldf l.lnlgoldf l.lnwtioil
- test　l.lnwtioil
- reg　lnwtioil l.lnwtioil l.lnlgoldf
- test　l.lnlgoldf

03 设置完毕后，等待输出结果。

20.7.2 结果分析

在 Stata 14.0 主界面的结果窗口我们可以看到如图 20.34~图 20.35 所示的分析结果。

```
. reg  lnlgoldf l.lnlgoldf l.lnwtioil

      Source |       SS           df       MS      Number of obs   =        48
-------------+----------------------------------   F(  2,      45)  =    514.93
       Model |  1.01745187         2  .508725934   Prob > F        =    0.0000
    Residual |  .044457654        45  .000987948   R-squared       =    0.9581
-------------+----------------------------------   Adj R-squared   =    0.9563
       Total |  1.06190952        47   .02259382   Root MSE        =    .03143

------------------------------------------------------------------------------
    lnlgoldf |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lnlgoldf |
         L1. |   .9530761   .0740646    12.87   0.000     .8039023     1.10225
             |
    lnwtioil |
         L1. |   .0241425   .0344262     0.70   0.487    -.0451954     .0934803
             |
       _cons |   .2052832    .331026     0.62   0.538    -.4614374     .8720037
------------------------------------------------------------------------------

. test  l.lnwtioil

 ( 1)  L.lnwtioil = 0

       F(  1,    45) =    0.49
            Prob > F =    0.4867
```

图 20.34　分析结果图 1

图 20.34 展示的是原油价格是否是黄金价格的格兰杰因的检验结果。通过观察分析结果可以看出 l.lnwtioil 的系数值是非常不显著的。具体体现在其 t 值、F 值以及 P 值上，关于这一结果的详细解读方法前面章节中多有提及，限于篇幅此处不再赘述，所以可以比较有把握地得出结论，原油价格不是黄金价格的格兰杰因。

图 20.35 展示的是黄金价格是否是原油价格的格兰杰因的检验结果。通过观察分析结果可以看出 l.lnlgoldf 的系数值是不显著的。具体体现在其 t 值、F 值以及 P 值上，关于这一结果的详细解读方法前面章节中多有提及，限于篇幅此处不再赘述。但是，我们在前面章节中曾经提到存在协整关系的变量间至少有一种格兰杰因果关系，所以可以相对地认为黄金价格是原油价格的格兰杰因。

```
. reg  lnwtioil l.lnwtioil l.lnlgoldf

      Source |       SS           df       MS      Number of obs   =        48
-------------+----------------------------------   F(  2,      45)  =    397.55
       Model |  4.36148081         2   2.1807404   Prob > F        =    0.0000
    Residual |  .246847782        45  .005485506   R-squared       =    0.9464
-------------+----------------------------------   Adj R-squared   =    0.9441
       Total |  4.60832859        47  .098049544   Root MSE        =    .07406

------------------------------------------------------------------------------
    lnwtioil |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lnwtioil |
         L1. |   .8482087   .0811204    10.46   0.000     .6848239    1.011594
             |
    lnlgoldf |
         L1. |   .2683232   .1745227     1.54   0.131    -.0831836      .61983
             |
       _cons |  -1.019525   .7800159    -1.31   0.198    -2.590558    .5515074
------------------------------------------------------------------------------

. test  l.lnlgoldf

 ( 1)  L.lnlgoldf = 0

       F(  1,    45) =    2.36
            Prob > F =    0.1312
```

图 20.35　分析结果图 2

20.8 建立模型

在经过了对数据进行描述性分析、绘制变量时间序列趋势图简要分析数据特征、进行相关性检验探索变量之间的相关关系、进行单位根检验综合分析数据平稳性、使用协整检验方式分析数据长期均衡关系、进行格兰杰因果关系检验探讨变量因果关系之后，本节进行最后的步骤，就是根据前面得出的一系列结论建立相应的数据模型。建立模型的步骤如下。

1. 建立模型方程

根据前面几节的分析构建如下所示的模型方程：

$$d.lnwtioil = a + b*dl.lnwtioil + c*d.lnlgoldf + d*ecm_{t-1} + u$$

其中，a、b、c、d 为系数，ecm 为误差修正项，u 为误差扰动项。

ecm 误差修正项的模型方程为：

$$ecm_t = lnwtioil - a*lnlgoldf - b$$

其中，a、b 为系数。实质上，ecm 是该模型方程的误差扰动项，或者说以 lnwtioil 为因变量，以 lnlgoldf 为自变量进行最小二乘估计回归后的残差。

2. 估计残差序列

在主界面的“Command”文本框中输入命令：

```
reg  lnwtioil lnlgoldf
predict e,resid
```

并按键盘上的回车键分别进行确认，即可出现如图 20.36 所示的残差序列。

	month	lgoldf	wtioil	lnlgoldf	lnwtioil	e
1	1	281.65	19.67	5.640665	2.979095	-.0682274
2	2	295.5	20.74	5.688669	3.032064	-.1065131
3	3	294.05	24.42	5.68375	3.195402	.0661762
4	4	302.68	26.27	5.712676	3.268428	.0842123
5	5	314.49	27.02	5.750952	3.296577	.0395989
6	6	310.25	25.52	5.737379	3.239462	.0082874
7	7	313.29	26.94	5.747129	3.293612	.0439009
8	8	310.25	28.38	5.737379	3.345685	.1145096
9	9	319.16	29.67	5.765693	3.390136	.1051363
10	10	316.56	28.85	5.757513	3.36211	.0926602
11	11	319.15	26.27	5.765661	3.268428	-.0165128
12	12	332.43	29.42	5.806429	3.381675	.019234
13	13	356.86	32.94	5.877344	3.494688	-.0025613
14	14	359.32	35.87	5.884213	3.579901	.0695926
15	15	340.55	33.55	5.830562	3.513037	.1047195
16	16	328.58	28.25	5.79478	3.341094	.0007978
17	17	355.68	28.14	5.874032	3.337192	-.1537607
18	18	356.53	30.72	5.876419	3.424914	-.0705766
19	19	351	30.76	5.860786	3.426215	-.0395586
20	20	359.77	31.59	5.885465	3.452841	-.0598475
21	21	378.95	28.29	5.937404	3.342508	-.268916
22	22	378.92	30.33	5.937325	3.412137	-.1991366
23	23	389.91	31.09	5.965916	3.436886	-.2287394
24	24	407.59	32.15	6.010262	3.470412	-.279514
25	25	413.99	34.27	6.025842	3.53427	-.2452742
26	26	405.33	34.74	6.004702	3.547892	-.1914652
27	27	406.67	36.76	6.008002	3.60441	-.1412214
28	28	403.02	36.69	5.998986	3.602504	-.1259878
29	29	383.4	40.28	5.949079	3.695855	.0622368

图 20.36 查看数据

3．估计误差修正项方程

在主界面的“Command”文本框中输入命令：

```
reg e lnwtioil lnlgoldf
```

并按键盘上的回车键进行确认，即可出现如图 20.37 所示的 ecm 误差修正项的模型方程估计结果。

```
. reg e lnwtioil lnlgoldf

      Source |       SS       df       MS              Number of obs =      49
-------------+------------------------------           F(  2,    46) =       .
       Model |  .855323994     2  .427661997           Prob > F      =       .
    Residual |           0    46           0           R-squared     =  1.0000
-------------+------------------------------           Adj R-squared =  1.0000
       Total |  .855323994    48  .01781925            Root MSE      =       0

------------------------------------------------------------------------------
           e |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lnwtioil |          1          .        .       .            .           .
    lnlgoldf |  -1.901004          .        .       .            .           .
       _cons |   7.675604          .        .       .            .           .
------------------------------------------------------------------------------
```

图 20.37　分析结果图 1

观察分析结果，我们得到的 ecm 模型方程为：

e=lnwtioil-1.901004*lnlgoldf+7.675604

该方程反映的是变量的长期均衡关系。

4．估计模型整体方程

在主界面的“Command”文本框中输入命令：

```
reg  d.lnwtioil dl.lnwtioil d.lnlgoldf l.e
```

并按键盘上的回车键进行确认，即可出现如图 20.38 所示的模型整体方程估计结果。

```
. reg  d.lnwtioil dl.lnwtioil d.lnlgoldf l.e

      Source |       SS       df       MS              Number of obs =      47
-------------+------------------------------           F(  3,    43) =    2.31
       Model |  .036936832     3  .012312277           Prob > F      =  0.0897
    Residual |  .229113878    43   .00532823           R-squared     =  0.1388
-------------+------------------------------           Adj R-squared =  0.0788
       Total |  .266050711    46  .005783711           Root MSE      =  .07299

------------------------------------------------------------------------------
  D.lnwtioil |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lnwtioil |
        LD.  |    .133125   .1472165     0.90   0.371    -.1637653    .4300154
             |
    lnlgoldf |
        D1.  |   .5852346   .3525989     1.66   0.104    -.1258489    1.296318
             |
           e |
        L1.  |  -.1847524   .0830589    -2.22   0.031    -.3522566   -.0172481
             |
       _cons |    .014413   .0121972     1.18   0.244     -.010185    .0390109
------------------------------------------------------------------------------
```

图 20.38　分析结果图 2

从上述分析结果中可以看到共有 47 个样本参与了分析。模型的 F 值(3, 43) = 2.31，P 值（Prob > F）= 0.0897，说明模型整体上还是可以接受的。模型的可决系数（R-squared）为 0.1388，模型修正的可决系数（Adj R-squared）为 0.0788，说明模型解释能力偏弱。

模型的回归方程是：

d.lnwtioil = 0.133125 *dl.lnwtioil+ 0.5852346*d1.lnlgoldf -0.1847524*l1.e+ 0.014413

变量 dl.lnwtioi 的系数标准误是 0.1472165，t 值为 0.90，P 值为 0.371，系数是非常不显著的，95%的置信区间为[-0.1637653，0.4300154]。变量 d1.lnlgoldf 的系数标准误是 0.3525989，t 值为 1.66，P 值为 0.104，系数也是非常不显著的，95%的置信区间为[-0.1258489 , 1.296318]。变量 l.e 的系数标准误是 0.0830589，t 值为-2.22，P 值为 0.031，系数是比较显著的，95%的置信区间为[-0.3522566 , -0.0172481]。常数项的系数标准误是 0.0121972，t 值为 1.18，P 值为 0.244，系数也是非常不显著的，95%的置信区间为[-0.010185 , 0.0390109]。

从上面的分析中可以看出，变量间的短期关系是比较不显著的，但是变量的长期均衡关系却很显著。

20.9　研究结论

经过前面的研究之后，可以比较有把握地得出以下研究结论：

- 黄金价格和原油价格都不是平稳的，都是具有长期增长趋势，并且存在很多波动的。我们从时间序列走势图上可以看出两个变量的长期增长性，从单位根检验结果上可以看出黄金价格和原油价格的不平稳性。
- 黄金价格和原油价格之间是存在长期均衡关系的，这一点可以从协整检验的结论上看出来。这意味着黄金价格和原油价格存在某种价格联动关系，在长期中是可以找到变动规律的。
- 原油价格不是黄金价格的格兰杰因，但黄金价格是原油价格的格兰杰因。或者说，黄金价格的变动会引起原油价格的相应变动，但原油价格的变动未必会引起黄金价格的相应变动。
- 黄金价格和原油价格长期是一种正向变动关系。这一点从误差修正项方程上就能看出来，得出的误差修正项方程是 e=lnwtioil-1.901004*lnlgoldf+7.675604，在探讨长期关系时，e 取值为 0，那么方程就变为 lnwtioil=1.901004*lnlgoldf-7.675604，所以黄金价格的变化会引起原油价格的同向变化，当黄金价格升高时，原油价格会随之升高。
- 短期内，黄金价格和原油价格也是一种正向变动关系，但是这种短期关系远远不如长期关系更明显。一方面体现在黄金价格作为自变量的系数值上，在长期关系下系数值更大；另一方面体现在变量的显著性上，在长期关系下系数的显著程度更高。
- 长期均衡关系的存在可以较好地平抑短期波动。我们可以看到模型整体的回归方程中误差修正项的系数是负值而且非常显著，这说明长期均衡关系可以有效削弱短期内变量的剧烈波动。例如黄金价格突然大幅度迅速上涨，那么由于模型中长期关系的存在，误差修正项也会随着提升，从而使短期内原油价格不会提升太多。

20.10 本章习题

表 20.2 给出了某企业经营资产和经营利润的有关数据，试使用描述性分析、时间序列趋势图分析、相关性检验、单位根检验、协整检验、格兰杰因果关系检验等方法研究数据特征并对变量间的关系进行分析，最后建立相应的方程模型描述两者之间的联动关系。

表 20.2 某企业经营资产和经营利润的有关数据

月份	经营资产/万元	经营利润/万元
1	283.9	22.89
2	286.9	23.15
3	291.5	24.12
4	303.33	25.19
5	314.49	27.02
6	310.25	25.52
…	…	…
45	456.05	66.32
46	470.3	63.12
47	472.69	59.89
48	512.9	58.49
49	550.96	67.79

第 21 章　Stata 在 ROE 与股权集中度之间关系研究中的应用

企业管理者总是希望能探寻到最佳的组织架构，以便在资源既定的前提下实现企业的最优化经营，所以企业经营业绩和股权集中度之间的关系历来是学者们研究的热点。本章选取在沪深两市上市的我国 14 家上市银行在 2008 年前三季度的数据作为样本进行了观测，并使用 Stata 14.0 对数据进行了深入分析，发现我国上市银行的净资产收益率与其第一大股东的持股量之间存在着倒“U”型关系。

21.1　研究背景

关于股权集中度问题的研究起源于 Berle 与 Means（1933），他们认为随着所有权的扩散，现代公司中典型的股东已不再能真正行使有效监督经营者行为的权利，而所有者与经营者的利益冲突的结果总是以有利于经营者一方而结束，私人财产的社会功能也因此受到严重的损害。Jensen 和 Meckling（1976）对公司价值与经理所拥有的股权之间的关系进行了研究，认为公司价值取决于内部股东所占有股份的比例，这一比例越高，公司的价值就越大。

其后，国外的相关研究主要集中在“股权集中度与企业经营业绩和企业市场价值是否存在显著的相关关系”方面，但是并无明确一致的实证结果。Demsets 和 Lehn（1985）考察了《财富》上 511 家美国大公司，发现股权集中度与 ROE 并不相关。Shlelter 和 Vishny（1986、1997）认为大股东但不控股股东的存在有利于改善公司的控制问题，进而增加公司价值。Mcconnell Servaes（1990）认为公司价值是公司股权结构的函数，他们通过对 1986 年 1093 个样本公司的市场价值与公司资产重置价值的比值和股权结构关系的实证分析，得出一个具有显著性的结论，即此比值与股权之间具有曲线关系，股权从 0 增加至 40%时，曲线向上倾斜，比例达到 40%~50%时，曲线开始向下倾斜。Mehran（1995）研究发现股权集中度与企业的 TobinQ 值、ROE 均无显著相关关系。Han 和 Suk（1998）研究发现，公司业绩与外部大量持股股东的股权比例呈正相关。

国内关于股权集中度的研究文献主要有：许小年（1997）的研究表明国有股比重大，公司效益差，而法人股则相反。陈晓和江东（2000）引入行业变量，发现公司业绩与股权结构相关，但股权多元化发挥功能的前提是提高行业竞争性。陈小悦和徐晓东（2001）在划分保护性和非保护性行业后，发现在非保护性行业第一大股东持股比例与业绩正相关，国有股和法人股比例与业绩关系不显著。朱武祥和宋勇（2001）重点以家电行业为样本论证了股权结构与公司业绩之间并不存在显著关系。

21.2 基本概念与数据说明

股权集中度（Concentration Ratio of Shares）是指全部股东因持股比例的不同所表现出来的股权集中还是分散的数量化指标，是衡量公司的股权分布状态的主要指标，也是衡量公司稳定性强弱的重要指标。

本章采用的是第一大股东的持股量、前五大股东的持股量、前十大股东的持股量以及它们各自的平方项。

公司绩效是指公司经营的业绩和效率，它反映公司的经营效果，一般用某个或一组财务指标加以反映，目前国内外股权结构研究一般采用托宾 Q 比率、净资产收益率（ROE）及主营业务资产收益率（CROA）作为公司绩效的评价标准。

- 托宾 Q 比率：Q=企业市场价值/企业重置成本=（权益市场总值+负债总值）/公司总产账面价值。
- 净资产收益率：ROE =净利润/净资产。
- 主营业务资产收益率：CROA＝ 主营业务利润/总资产。

本章采用的是 ROE 指标。ROE 指标反映了一定资本量下的相对利润水平，体现了资产的盈利能力，是资产是否优良的重要衡量指标。尽管更严格意义上的定义应该是将其中非主营利润从公司盈利中剔除，甚至还应该对公司的净资产指标进行严格评估，但是就整体统计层面上，ROE 水平应该是一个非常好的指标。

受前人研究的启发，本章选取了在沪深两市上市的中国 14 家上市银行在 2008 年前三季度的数据作为样本，进行了观测，发现我国上市银行的净资产收益率与其第一大股东的持股量之间存在着倒“U”型关系。

样本数据为面板数据，上市银行包括深圳发展银行、宁波银行、浦发银行、华夏银行、民生银行、招商银行、南京银行、兴业银行、北京银行、交通银行、工商银行、建设银行、中国银行和中信银行。时间点分别为 2008 年 9 月 30 日、2008 年 6 月 30 日和 2008 年 3 月 31 日。数据来源于中国上市公司资讯网，其中 ROE 的数据和前十大股东的各自数据可以直接从网上得到。因为本章还试图以前五大股东或者前十大股东的总持股量作为解释变量，所以手工计算了前五大股东或者前十大股东的总持股量（具体数据见表 21.1）。

表 21.1 沪深两市上市的中国 14 家上市银行在 2008 年前三季度的数据

上市银行	第一大股东的持股量（比例）	前五大股东的持股量（比例）	前十大股东的持股量（比例）	净资产收益率	时间（1代表20080930，2代表20080630，3代表20080331）
深发A	6.55	19.43	26.95	18.05	1
深发A	16.76	27.54	34.03	12.65	2
深发A	1.62	7.62	13.86	7.15	3
宁波银行	2.45	8.52	9.99	13.1	1
宁波银行	10.8	42.28	62.07	8.89	2
宁波银行	0.29	0.74	1.03	4.06	3
…	…	…	…	…	…

（续表）

上市银行	第一大股东的持股量（比例）	前五大股东的持股量（比例）	前十大股东的持股量（比例）	净资产收益率	时间（1代表20080930，2代表20080630，3代表20080331）
中国银行	67.49	95.45	97.25	13.03	1
中国银行	67.49	95.46	97.26	9.65	2
中国银行	67.49	95.3	97.24	4.86	3
中信银行	62.33	94.73	95.33	13.19	1
中信银行	62.33	94.74	95.34	9.33	2
中信银行	62.33	94.74	95.34	4.86	3

21.3 实证分析

	下载资源:\video\chap21\…
	下载资源:\sample\chap21\案例21.dta

21.3.1 描述性分析

在用 Stata 进行分析之前，我们要把数据录入到 Stata 中。本例中有 9 个变量，分别为第一大股东的持股量、前五大股东的持股量、前十大股东的持股量、净资产收益率、时间、第一大股东的持股量的平方、前五大股东的持股量的平方、前十大股东的持股量的平方、银行名称。我们把第一大股东的持股量变量设定为 top1，把前五大股东的持股量变量设定为 top5，把前十大股东的持股量变量设定为 top10，把净资产收益率变量设定为 roe，把时间变量设定为 t，把第一大股东的持股量的平方变量设定为 stop1，把前五大股东的持股量的平方变量设定为 stop5，把前十大股东的持股量的平方变量设定为 stop10，把银行名称变量设定为 bank。变量类型及长度采取系统默认方式，然后录入相关数据。相关操作我们在第 1 章中已有详细讲述。录入完成后数据如图 21.1 所示。

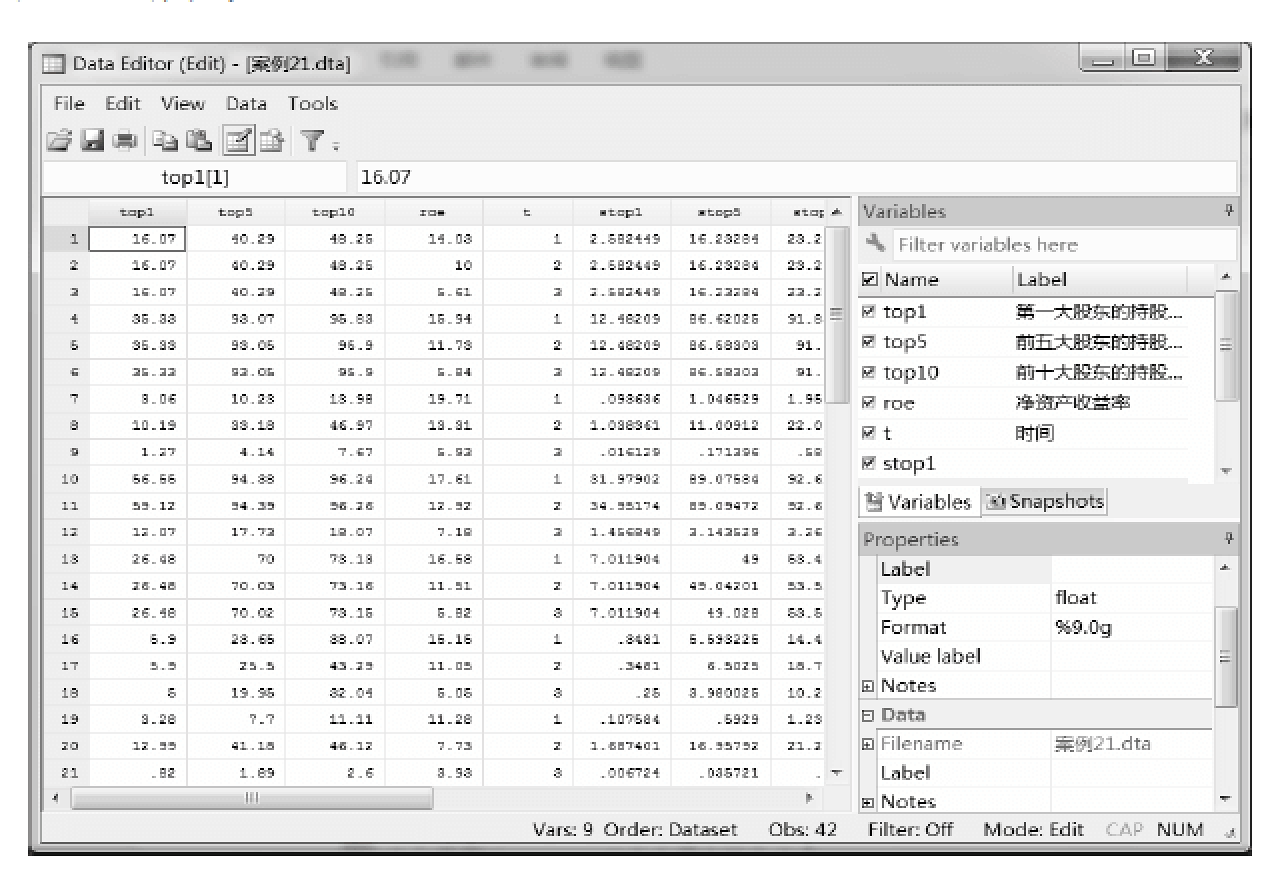


图 21.1 案例 21 数据

先做一下数据保存，然后开始展开分析，分析步骤及结果如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入命令：

```
summarize   top1 top5 top10 roe t stop1 stop5 stop10 bank
```

03 设置完毕后，按键盘上的回车键进行确认。

在 Stata 14.0 主界面的结果窗口可以看到如图 21.2 所示的分析结果。

```
. summarize   top1 top5 top10 roe t stop1 stop5 stop10 bank

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        top1 |        42    23.61881    21.56027        .29      67.49
        top5 |        42    47.22786    32.81483        .74      95.46
       top10 |        42    53.05762    31.36387       1.03      97.26
         roe |        42    11.68238    5.407001       3.93      25.84
           t |        42           2    .8263939          1          3
-------------+--------------------------------------------------------
       stop1 |        42    10.11626    15.19319    .000841     45.549
       stop5 |        42    32.81645    35.96414    .005476   91.12611
      stop10 |        42    37.75382    35.62448    .010609   94.59508
        bank |        42         7.5    4.079993          1         14

.
```

图 21.2　描述性分析结果图

通过观察分析结果，可以对沪深两市上市的中国 14 家上市银行在 2008 年前三季度的数据有整体初步的了解。从结果可以看出，有效观测样本共有 42 个。第一大股东的持股量的均值是 23.61881，标准差是 21.56027，最小值是 0.29，最大值是 67.49；前五大股东的持股量的均值是 47.22786，标准差是 32.81483，最小值是 0.74，最大值是 95.46；前十大股东的持股量的均值是 53.05762，标准差是 31.36387，最小值是 1.03，最大值是 97.26；净资产收益率的均值是 11.68238，标准差是 5.407001，最小值是 3.93，最大值是 25.84；此处时间变量被简单地看成了定距变量，按定距变量的描述性统计进行了处理，均值是 2，标准差是 0.8263939，最小值是 1，最大值是 3；第一大股东的持股量的平方的均值是 10.11626，标准差是 15.19319，最小值是 0.000841，最大值是 45.549；前五大股东的持股量平方的均值是 32.81645，标准差是 35.96414，最小值是 0.005476，最大值是 91.12611；前十大股东的持股量的平方的均值是 37.75382，标准差是 35.62448，最小值是 0.010609，最大值是 94.59508；此处时间变量被简单地看成了定距变量，其最小值为 1，最大值为 14，说明共有 14 家银行参与了分析过程。

21.3.2　图形分析

图形分析步骤及结果如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入命令：

- twoway scatter roe top1：本命令旨在绘制净资产收益率和第一大股东的持股量的散点图。

- twoway scatter roe top5：本命令旨在绘制净资产收益率和前五大股东的持股量的散点图。
- twoway scatter roe top10：本命令旨在绘制净资产收益率和前十大股东的持股量的散点图。
- twoway scatter roe stop1：本命令旨在绘制净资产收益率和第一大股东的持股量的平方的散点图。
- twoway scatter roe stop5：本命令旨在绘制净资产收益率和前五大股东的持股量的平方的散点图。
- twoway scatter roe stop10：本命令旨在绘制净资产收益率和前十大股东的持股量的平方的散点图。

03 设置完毕后，按键盘上的回车键进行确认。

在 Stata 14.0 主界面的结果窗口可以看到如图 21.3~图 21.8 所示的分析结果。

图 21.3 是净资产收益率和第一大股东的持股量的散点图。

图 21.4 是净资产收益率和前五大股东的持股量的散点图。

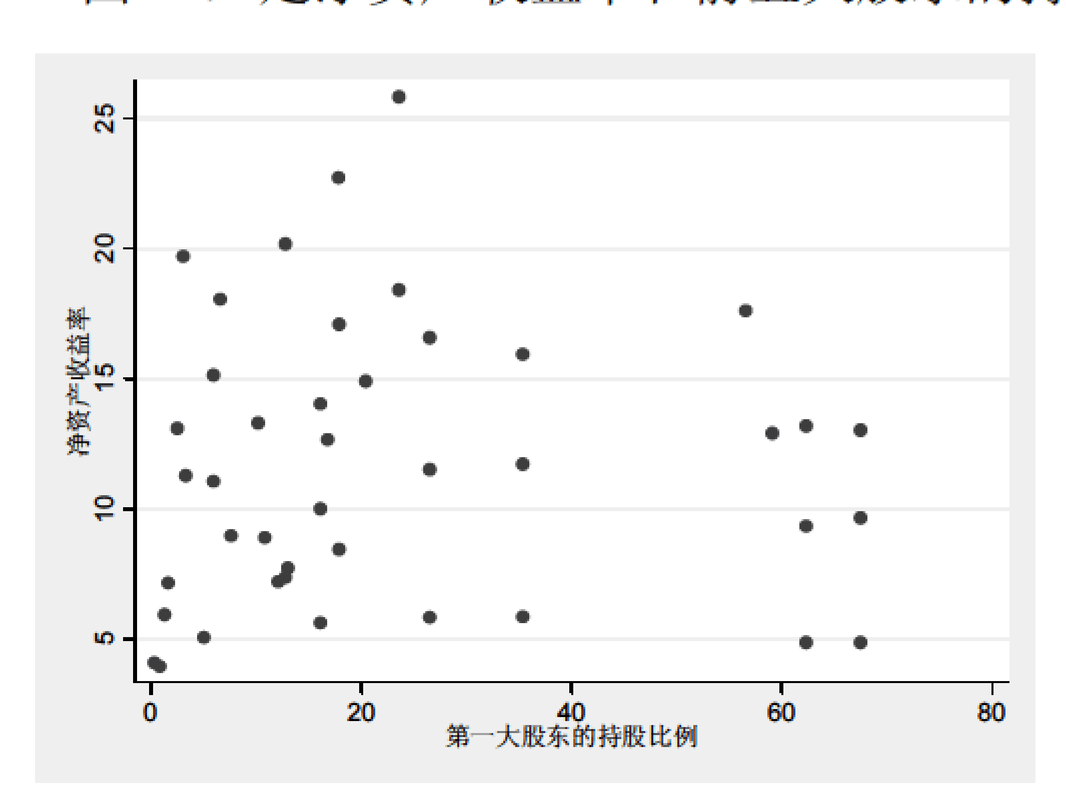


图 21.3　图形分析结果 1

图 21.4　图形分析结果 2

图 21.5 是净资产收益率和前十大股东的持股量的散点图。

图 21.6 是净资产收益率和第一大股东的持股量的平方的散点图。

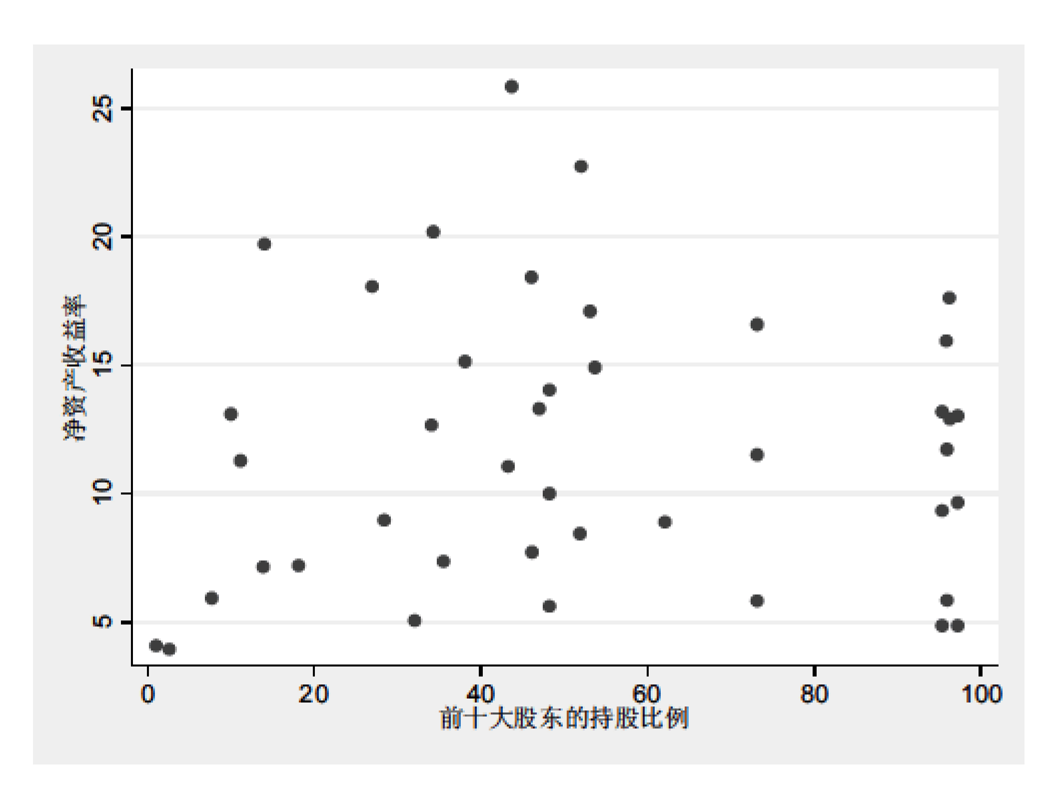


图 21.5　图形分析结果 3

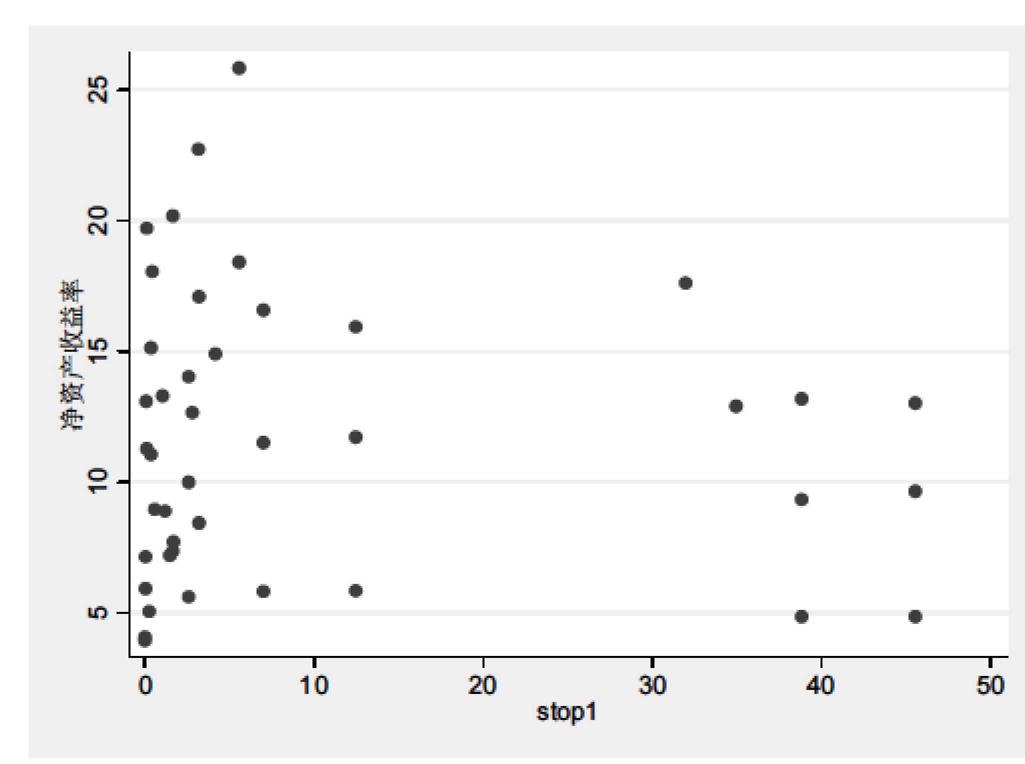


图 21.6　图形分析结果 4

图 21.7 是净资产收益率和前五大股东的持股量的平方的散点图。

图 21.8 是净资产收益率和前十大股东的持股量的平方的散点图。

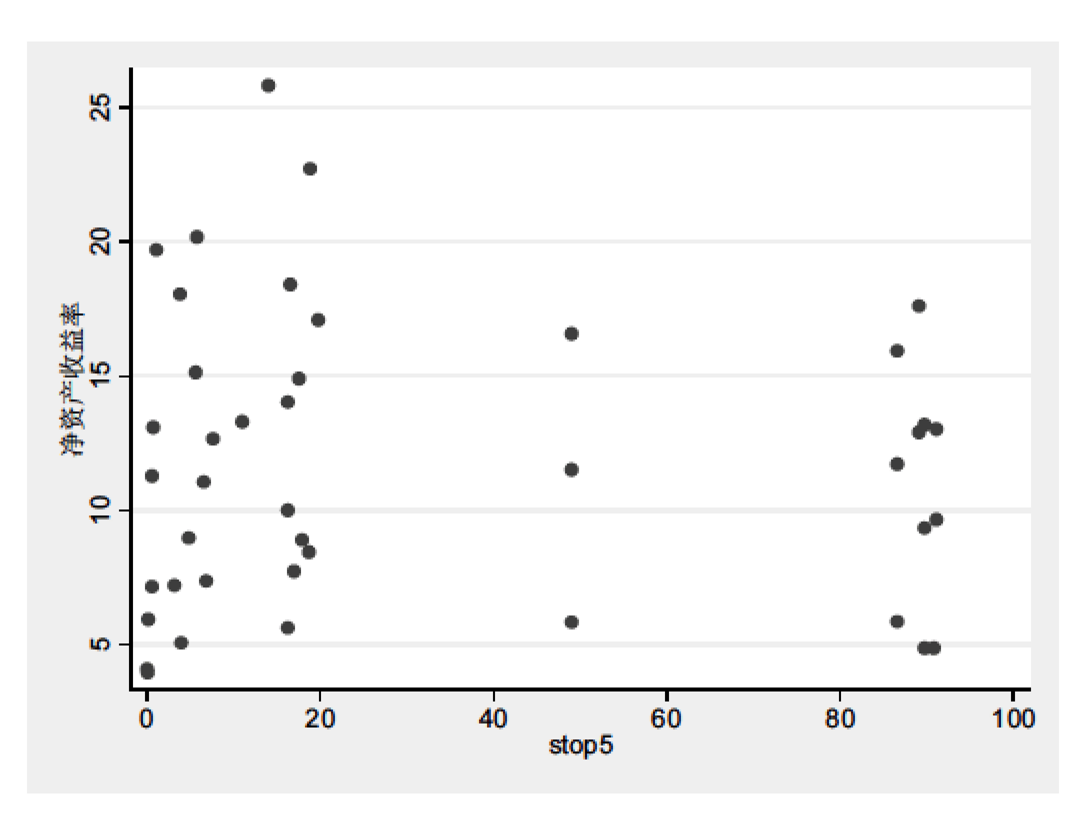


图 21.7 图形分析结果 5

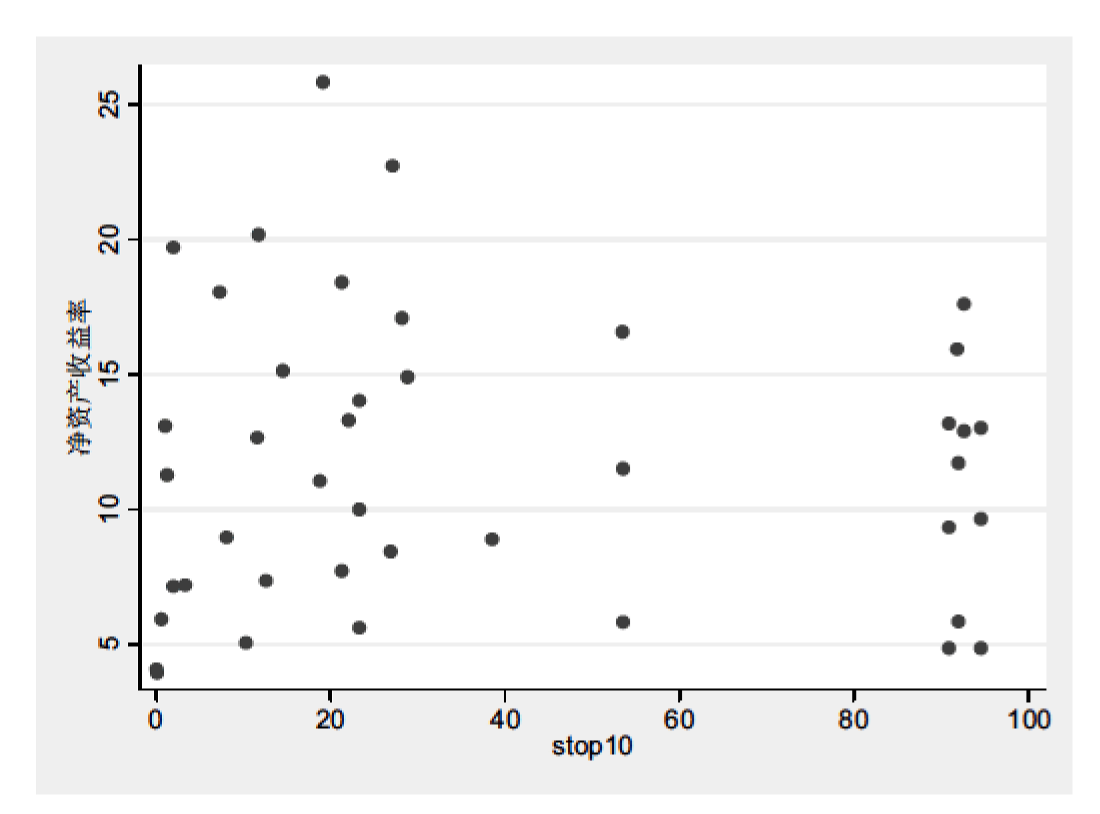


图 21.8 图形分析结果 6

从图上可以发现我国上市银行的净资产收益率与股权集中度之间似乎并没有显著的关系。

21.3.3 普通最小二乘回归分析

下面以 ROE 为被解释变量，以第一大股东持股量（top1）、前五大股东持股量（top5）、前十大股东持股量（top10）、第一大股东持股量的平方除以 100（stop1）、前五大股东持股量的平方除以 100（stop5）、前十大股东持股量的平方除以 100（stop10）为解释变量，后 3 项之所以除以 100 是为了使解释变量数据之间的差距不致于过大。

建立线性模型：

ROE=a*top1+b*top5+c*top10+d*stop1+e*stop5+f*stop10+u

普通最小二乘回归分析的步骤及结果如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入如下命令。

- sw regress roe top1 top5 top10 stop1 stop5 stop10,pr(0.05)：本命令的含义是使用逐步回归分析方法，以净资产收益率为因变量，以第一大股东的持股量、前五大股东的持股量、前十大股东的持股量、第一大股东的持股量的平方、前五大股东的持股量的平方、前十大股东的持股量的平方等变量为自变量，进行最小二乘回归分析。
- reg roe top1 top5 top10 stop1 stop5 stop10,vce(cluster bank)：本命令的含义是以净资产收益率为因变量，以第一大股东的持股量、前五大股东的持股量、前十大股东的持股量、第一大股东的持股量的平方、前五大股东的持股量的平方、前十大股东的持股量的平方等变量为自变量，并使用以“bank”为聚类变量的聚类稳健标准差，进行最小二乘回归分析。
- reg roe top1 top5 stop1,vce(cluster bank)：本命令是在上步回归的基础上，剔除掉不显著的自变量以后，以净资产收益率为因变量，以第一大股东的持股量、前五大股东的持股量、第一大股东的持股量的平方等变量为自变量，并使用以“bank”为聚类变量的聚类稳健标准差，进行最小二乘回归分析。

03 设置完毕后，按键盘上的回车键进行确认。

在 Stata 14.0 主界面的结果窗口可以看到如图 21.9~图 21.11 所示的分析结果。

图 21.9 是使用逐步回归分析方法，以净资产收益率为因变量，以第一大股东的持股量、前五大股东的持股量、前十大股东的持股量、第一大股东的持股量的平方、前五大股东的持股量的平方、前十大股东的持股量的平方等变量为自变量，进行最小二乘回归分析的结果。

```
. sw regress roe top1 top5 top10 stop1 stop5 stop10,pr(0.05)
                      begin with full model
p = 0.4778 >= 0.0500  removing stop10
p = 0.5920 >= 0.0500  removing stop5
p = 0.2445 >= 0.0500  removing top10

      Source |       SS           df       MS      Number of obs   =        42
-------------+----------------------------------   F(3, 38)        =      3.28
       Model |  246.501432         3  82.1671441   Prob > F        =    0.0312
    Residual |  952.160703        38  25.0568606   R-squared       =    0.2056
-------------+----------------------------------   Adj R-squared   =    0.1429
       Total |  1198.66214        41  29.2356618   Root MSE        =    5.0057

------------------------------------------------------------------------------
         roe |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        top1 |   .9265336   .3214238     2.88   0.006     .2758452    1.577222
        top5 |  -.1944912   .0941327    -2.07   0.046     -.385053   -.0039295
       stop1 |  -.9541558   .3058385    -3.12   0.003    -1.573294   -.3350181
       _cons |    8.63665   1.701976     5.07   0.000     5.191179    12.08212
------------------------------------------------------------------------------
```

图 21.9　普通最小二乘回归分析结果 1

从上述分析结果中可以看出共有 42 个样本参与了分析，模型的 F 值(3, 38) =3.28，P 值（Prob > F）= 0.0312，说明模型整体上是非常显著的。模型的可决系数（R-squared）为 0.2056，模型修正的可决系数（Adj R-squared）为 0.1429，说明模型的解释能力还是差强人意的。

变量 top1 的系数标准误是 0.3214238，t 值为 2.88，P 值为 0.006，系数是非常显著的，95%的置信区间为[0.2758452, 1.577222]。变量 top5 的系数标准误是 0.0941327，t 值为-2.07，P 值为 0.046，系数是非常显著的，95%的置信区间为[-0.385053, -0.0039295]。变量 stop1 的系数标准误是 0.3058385，t 值为-3.12，P 值为 0.003，系数是非常显著的，95%的置信区间为[-1.573294, -0.3350181]。常数项的系数标准误是 1.701976，t 值为 5.07，P 值为 0.000，系数也是非常显著的，95%的置信区间为[5.191179, 12.08212]。

模型的回归方程是：

ROE=0.9265336*top1-0.9541558*stop1-0.1944912*top5+8.63665

可以看出 stop1 前面的系数显著为负，说明中国上市银行的 ROE 与第一大股东持股量之间显著存在着倒“U”型关系。

图 21.10 是以净资产收益率为因变量，以第一大股东的持股量、前五大股东的持股量、前十大股东的持股量、第一大股东的持股量的平方、前五大股东的持股量的平方、前十大股东的持股量的平方等变量为自变量，并使用以“bank”为聚类变量的聚类稳健标准差，进行最小二乘回归分析的结果。

```
. reg  roe top1 top5 top10 stop1 stop5 stop10,vce(cluster  bank)

Linear regression                               Number of obs =        42
                                                F(  6,     13) =      8.38
                                                Prob > F       =    0.0007
                                                R-squared      =    0.2518
                                                Root MSE       =    5.0622

                                  (Std. Err. adjusted for 14 clusters in bank)
```

roe	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
top1	1.126628	.4611476	2.44	0.030	.1303789	2.122877
top5	-.9885495	.4308575	-2.29	0.039	-1.919361	-.0577384
top10	.6892669	.3152673	2.19	0.048	.0081732	1.370361
stop1	-1.165315	.4522619	-2.58	0.023	-2.142367	-.1882624
stop5	.568966	.343857	1.65	0.122	-.173892	1.311824
stop10	-.4732053	.3197272	-1.48	0.163	-1.163934	.2175233
_cons	6.171474	1.431993	4.31	0.001	3.077841	9.265107

图 21.10　普通最小二乘回归分析结果 2

可以看出，使用以净资产收益率为因变量，以第一大股东的持股量、前五大股东的持股量、前十大股东的持股量、第一大股东的持股量的平方、前五大股东的持股量的平方、前十大股东的持股量的平方等变量为自变量，并使用以“bank”为聚类变量的聚类稳健标准差，进行最小二乘回归分析的结果较普通最小二乘回归分析在模型解释能力上有所提高。

图 21.11 是在上步回归的基础上，剔除不显著的自变量以后，以净资产收益率为因变量，以第一大股东的持股量、前五大股东的持股量、第一大股东的持股量的平方等变量为自变量，并使用以“bank”为聚类变量的聚类稳健标准差，进行最小二乘回归分析的结果。

可以看出，在剔除不显著的自变量以后，以净资产收益率为因变量，以第一大股东的持股量、前五大股东的持股量、第一大股东的持股量的平方等变量为自变量，并使用以“bank”为聚类变量的聚类稳健标准差，进行最小二乘回归分析的结果与普通最小二乘回归分析大同小异。

```
. reg  roe top1 top5 stop1,vce(cluster  bank)

Linear regression                               Number of obs =        42
                                                F(  3,     13) =      4.60
                                                Prob > F       =    0.0209
                                                R-squared      =    0.2056
                                                Root MSE       =    5.0057

                                  (Std. Err. adjusted for 14 clusters in bank)
```

roe	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
top1	.9265336	.3346691	2.77	0.016	.2035251	1.649542
top5	-.1944912	.0886654	-2.19	0.047	-.3860413	-.0029412
stop1	-.9541558	.3145051	-3.03	0.010	-1.633603	-.2747089
_cons	8.63665	1.236456	6.99	0.000	5.965448	11.30785

图 21.11　普通最小二乘回归分析结果 3

21.3.4　面板数据回归分析

下面以 ROE 为被解释变量，以第一大股东持股量（top1）、前五大股东持股量（top5）、前十大股东持股量（top10）、第一大股东持股量的平方除以 100（stop1）、前五大股东持股量的平方除以 100（stop5）、前十大股东持股量的平方除以 100（stop10）为解释变量，后三项之所以除以 100 是为了使解释变量数据之间的差距不致于过大。

建立线性模型：

ROE=a*top1+b*top5+c*top10+d*stop1+e*stop5+f*stop10+u

面板数据回归分析的步骤及结果如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入如下命令：

- list roe top1 top5 top10 stop1 stop5 stop10: 本命令的含义是对 7 个变量所包含的样本数据进行一一展示，以便简单直观地观测出数据的具体特征，为深入分析做好必要准备。
- xtset　bank t: 本命令的含义是对面板数据进行定义，其中横截面维度变量为我们上步生成的 bank，时间序列变量为 t。
- xtdes: 本命令旨在观测面板数据的结构，考察面板数据特征，为后续分析做好必要准备。
- xtsum: 本命令旨在显示面板数据组内、组间以及整体的统计指标。
- xttab　roe: 本命令旨在显示“roe”变量组内、组间以及整体的分布频率。
- xttab top1: 本命令旨在显示“top1”变量组内、组间以及整体的分布频率。
- xttab top5: 本命令旨在显示“top5”变量组内、组间以及整体的分布频率。
- xttab stop1: 本命令旨在显示“stop1”变量组内、组间以及整体的分布频率。
- xtline roe: 本命令旨在对每个个体显示“roe”变量的时间序列图。
- xtline top1: 本命令旨在对每个个体显示“top1”变量的时间序列图。
- xtline top5: 本命令旨在对每个个体显示“top5”变量的时间序列图。
- xtline stop1: 本命令旨在对每个个体显示“stop1”变量的时间序列图。
- xtreg roe top1 top5 stop1,fe vce(cluster　bank): 本命令的含义是以 roe 为因变量，以 top1、top5、stop1 为自变量，并使用以“bank”为聚类变量的聚类稳健标准差，进行固定效应回归分析。
- xtreg roe top1 top5 stop1,fe: 本命令的含义是以 roe 为因变量，以 top1、top5、stop1 为自变量，进行固定效应回归分析。
- estimates store fe: 本命令的含义是存储固定效应回归分析的估计结果。
- xi:xtreg roe top1 top5 stop1 i.bank,vce(cluster bank): 本命令旨在通过构建最小二乘虚拟变量模型来分析固定效应模型是否优于最小二乘回归分析。
- tab　t,gen(t): 本命令旨在创建年度变量的多个虚拟变量。
- xtreg roe top1 top5 stop1 t2-t3,fe vce(cluster bank): 本命令旨在通过构建双向固定效应模型来检验模型中是否应该包含时间效应。
- test　t2 t3: 本命令的含义是在上步回归的基础上，通过测试各虚拟变量的系数联合显著性来检验是否应该在模型中纳入时间效应。
- xtreg roe top1 top5 stop1,re vce(cluster bank): 本命令的含义是以 roe 为因变量，以 top1、top5、stop1 为自变量，并使用以“bank”为聚类变量的聚类稳健标准差，进行随机效应回归分析。
- xttest0: 本命令的含义是在上步回归的基础上，进行假设检验来判断随机效应模型是

否优于最小二乘回归模型。

- xtreg roe top1 top5 stop1,mle：本命令的含义是以 roe 为因变量，以 top1、top5、stop1 为自变量，并使用最大似然估计方法，进行随机效应回归分析。
- xtreg roe top1 top5 stop1,be：本命令的含义是以 roe 为因变量，以 top1、top5、stop1 为自变量，并使用组间估计量，进行组间估计量回归分析。
- xtreg roe top1 top5 stop1,re：本命令的含义是以 roe 为因变量，以 top1、top5、stop1 为自变量，进行随机效应回归分析。
- estimates store re：本命令的含义是存储随机效应回归分析的估计结果。
- hausman fe re,constant sigmamore：本命令的含义是进行豪斯曼检验，并据此判断应该选择固定效应模型还是随机效应模型。

03 设置完毕后，按键盘上的回车键进行确认。

在 Stata 14.0 主界面的结果窗口可以看到如图 21.12~图 21.37 所示的分析结果。

图 21.12 是对数据进行展示的结果。它的目的是通过对变量所包含的样本数据进行一一展示，以便简单直观地观测出数据的具体特征，为深入分析做好必要准备。

在如图 21.12 所示的分析结果中可以看出，数据的总体质量还是可以的，没有极端异常值，变量间的量纲差距也是可以接受的，可以进入下一步的分析。

```
. list roe top1 top5 top10 stop1 stop5 stop10
```

	roe	top1	top5	top10	stop1	stop5	stop10
1.	14.03	16.07	40.29	48.25	2.582449	16.23284	23.28062
2.	10	16.07	40.29	48.25	2.582449	16.23284	23.28062
3.	5.61	16.07	40.29	48.25	2.582449	16.23284	23.28062
4.	15.94	35.33	93.07	95.83	12.48209	86.62025	91.83389
5.	11.73	35.33	93.05	95.9	12.48209	86.58303	91.9681
6.	5.84	35.33	93.05	95.9	12.48209	86.58303	91.9681
7.	19.71	3.06	10.23	13.98	.093636	1.046529	1.954404
8.	13.31	10.19	33.18	46.97	1.038361	11.00912	22.06181
9.	5.93	1.27	4.14	7.67	.016129	.171396	.588289
10.	17.61	56.55	94.38	96.24	31.97902	89.07584	92.62137
11.	12.92	59.12	94.39	96.26	34.95174	89.09472	92.65988
12.	7.18	12.07	17.73	18.07	1.456849	3.143529	3.265249
13.	16.58	26.48	70	73.13	7.011904	49	53.47997
14.	11.51	26.48	70.03	73.16	7.011904	49.04201	53.52386
15.	5.82	26.48	70.02	73.15	7.011904	49.028	53.50923
16.	15.15	5.9	23.65	38.07	.3481	5.593225	14.49325
17.	11.05	5.9	25.5	43.29	.3481	6.5025	18.74024
18.	5.05	5	19.95	32.04	.25	3.980025	10.26562
19.	11.28	3.28	7.7	11.11	.107584	.5929	1.234321
20.	7.73	12.99	41.18	46.12	1.687401	16.95792	21.27054
21.	3.93	.82	1.89	2.6	.006724	.035721	.0676
22.	13.1	2.45	8.52	9.99	.060025	.7259041	.998001
23.	8.89	10.8	42.28	62.07	1.1664	17.87598	38.52685
24.	4.06	.29	.74	1.03	.000841	.005476	.010609
25.	25.84	23.57	37.4	43.7	5.555449	13.9876	19.0969
26.	18.42	23.57	40.63	46.09	5.555449	16.50797	21.24288
27.	8.95	7.61	21.89	28.37	.579121	4.791721	8.04857
28.	18.05	6.55	19.43	26.95	.429025	3.775249	7.263025
29.	12.65	16.76	27.54	34.03	2.808976	7.584517	11.58041
30.	7.15	1.62	7.62	13.86	.026244	.580644	1.920996
31.	20.19	12.78	24.03	34.24	1.633284	5.774409	11.72378
32.	14.91	20.4	41.91	53.65	4.1616	17.56448	28.78323
33.	7.35	12.78	26.14	35.48	1.633284	6.832996	12.5883
34.	22.73	17.8	43.37	52.02	3.1684	18.80957	27.0608
35.	17.09	17.88	44.42	53.07	3.196944	19.73136	28.16425
36.	8.45	17.88	43.22	51.87	3.196944	18.67969	26.90497
37.	13.03	67.49	95.45	97.25	45.549	91.10702	94.57562
38.	9.65	67.49	95.46	97.26	45.549	91.12611	94.59508
39.	4.86	67.49	95.3	97.24	45.549	90.82091	94.55618
40.	13.19	62.33	94.73	95.33	38.85029	89.73773	90.87809
41.	9.33	62.33	94.74	95.34	38.85029	89.75668	90.89715
42.	4.86	62.33	94.74	95.34	38.85029	89.75668	90.89715

图 21.12　面板数据回归分析结果 1

图 21.13 是对面板数据进行定义的结果，其中横截面维度变量为 bank，时间序列变量为 t。

```
. xtset  bank t
       panel variable:  bank (strongly balanced)
        time variable:  t, 1 to 3
                delta:  1 unit
```

图 21.13　面板数据回归分析结果 2

从图 21.13 中可以看出这是一个平衡的面板数据。

图 21.14 是面板数据结构的结果。

```
. xtdes

   bank:  1, 2, ..., 14                                       n =         14
      t:  1, 2, ..., 3                                        T =          3
          Delta(t) = 1 unit
          Span(t)  = 3 periods
          (bank*t uniquely identifies each observation)

Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                         3       3       3         3         3       3       3

     Freq.  Percent    Cum. |  Pattern
 ---------------------------+---------
       14    100.00  100.00 |  111
 ---------------------------+---------
       14    100.00         |  XXX
```

图 21.14　面板数据回归分析结果 3

从图 21.14 可以看出该面板数据的横截面维度 bank 为 1~14 共 14 个取值，时间序列维度 t 为 1~3 共 3 个取值，属于短面板数据，而且观测样本在时间上的分布也非常均匀。

图 21.15 是面板数据组内、组间以及整体的统计指标的结果。

在短面板数据中，同一时间段内的不同观测样本构成一个组。从图 21.15 中可以看出，变量 year 的组间标准差是 0，因为不同组的这一变量取值完全相同，同时变量 bank 的组内标准差也为 0，分布在同一组的数据属于同一个地区。

```
. xtsum

Variable         |      Mean   Std. Dev.       Min        Max |    Observations
-----------------+--------------------------------------------+----------------
top1     overall |  23.61881   21.56027        .29      67.49 |     N =      42
         between |             20.96504   4.513333      67.49 |     n =      14
         within  |             6.837636   -6.89119   40.15881 |     T =       3
                 |                                            |
top5     overall |  47.22786   32.81483        .74      95.46 |     N =      42
         between |             30.91719      15.85   95.40333 |     n =      14
         within  |             12.94482  -3.875475   72.78452 |     T =       3
                 |                                            |
top10    overall |  53.05762   31.36387       1.03      97.26 |     N =      42
         between |             28.39021   19.94333      97.25 |     n =      14
         within  |             14.73111    .937619   90.76429 |     T =       3
                 |                                            |
roe      overall |  11.68238   5.407001       3.93      25.84 |     N -      42
         between |             2.893701   7.646667   17.73667 |     n =      14
         within  |             4.612008   2.895714   19.78571 |     T =       3
                 |                                            |
t        overall |         2   .8263939          1          3 |     N =      42
         between |                    0          2          2 |     n =      14
         within  |             .8263939          1          3 |     T =       3
                 |                                            |
stop1    overall |  10.11626   15.19319    .000841     45.549 |     N =      42
         between |               14.977      .3154     45.549 |     n =      14
         within  |             4.179034  -11.22277   22.27213 |     T =       3
                 |                                            |
stop5    overall |  32.81645   35.96414    .005476   91.12611 |     N =      42
         between |             34.88239   3.980137   91.01801 |     n =      14
         within  |              11.6616  -24.47805   61.47314 |     T =       3
                 |                                            |
stop10   overall |  37.75382   35.62448    .010609   94.59508 |     N =      42
         between |             33.89342   6.921476   94.57563 |     n =      14
         within  |              13.2807  -21.82976   67.56487 |     T =       3
                 |                                            |
bank     overall |       7.5   4.079993          1         14 |     N =      42
         between |               4.1833          1         14 |     n -      14
         within  |                    0        7.5        7.5 |     T =       3
```

图 21.15　面板数据回归分析结果 4

图 21.16 是“roe”变量组内、组间以及整体的分布频率的结果。

. xttab roe

roe	Overall Freq.	Overall Percent	Between Freq.	Between Percent	Within Percent
3.93	1	2.38	1	7.14	33.33
4.06	1	2.38	1	7.14	33.33
4.86	2	4.76	2	14.29	33.33
5.05	1	2.38	1	7.14	33.33
5.61	1	2.38	1	7.14	33.33
5.82	1	2.38	1	7.14	33.33
5.84	1	2.38	1	7.14	33.33
5.93	1	2.38	1	7.14	33.33
7.15	1	2.38	1	7.14	33.33
7.18	1	2.38	1	7.14	33.33
7.35	1	2.38	1	7.14	33.33
7.73	1	2.38	1	7.14	33.33
8.45	1	2.38	1	7.14	33.33
8.89	1	2.38	1	7.14	33.33
8.95	1	2.38	1	7.14	33.33
9.33	1	2.38	1	7.14	33.33
9.65	1	2.38	1	7.14	33.33
10	1	2.38	1	7.14	33.33
11.05	1	2.38	1	7.14	33.33
11.28	1	2.38	1	7.14	33.33
11.51	1	2.38	1	7.14	33.33
11.73	1	2.38	1	7.14	33.33
12.65	1	2.38	1	7.14	33.33
12.92	1	2.38	1	7.14	33.33
13.03	1	2.38	1	7.14	33.33
13.1	1	2.38	1	7.14	33.33
13.19	1	2.38	1	7.14	33.33
13.31	1	2.38	1	7.14	33.33
14.03	1	2.38	1	7.14	33.33
14.91	1	2.38	1	7.14	33.33
15.15	1	2.38	1	7.14	33.33
15.94	1	2.38	1	7.14	33.33
16.58	1	2.38	1	7.14	33.33
17.09	1	2.38	1	7.14	33.33
17.61	1	2.38	1	7.14	33.33
18.05	1	2.38	1	7.14	33.33
18.42	1	2.38	1	7.14	33.33
19.71	1	2.38	1	7.14	33.33
20.19	1	2.38	1	7.14	33.33
22.73	1	2.38	1	7.14	33.33
25.84	1	2.38	1	7.14	33.33
Total	42	100.00	42	300.00	33.33
			(n = 14)		

图 21.16　面板数据回归分析结果 5

图 21.17 是“top1”变量组内、组间以及整体的分布频率的结果。

. xttab top1

top1	Overall Freq.	Overall Percent	Between Freq.	Between Percent	Within Percent
.29	1	2.38	1	7.14	33.33
.82	1	2.38	1	7.14	33.33
1.27	1	2.38	1	7.14	33.33
1.62	1	2.38	1	7.14	33.33
2.45	1	2.38	1	7.14	33.33
3.06	1	2.38	1	7.14	33.33
3.28	1	2.38	1	7.14	33.33
5	1	2.38	1	7.14	33.33
5.9	2	4.76	1	7.14	66.67
6.55	1	2.38	1	7.14	33.33
7.61	1	2.38	1	7.14	33.33
10.19	1	2.38	1	7.14	33.33
10.8	1	2.38	1	7.14	33.33
12.07	1	2.38	1	7.14	33.33
12.78	2	4.76	1	7.14	66.67
12.99	1	2.38	1	7.14	33.33
16.07	3	7.14	1	7.14	100.00
16.76	1	2.38	1	7.14	33.33
17.8	1	2.38	1	7.14	33.33
17.88	2	4.76	1	7.14	66.67
20.4	1	2.38	1	7.14	33.33
23.57	2	4.76	1	7.14	66.67
26.48	3	7.14	1	7.14	100.00
35.33	3	7.14	1	7.14	100.00
56.55	1	2.38	1	7.14	33.33
59.12	1	2.38	1	7.14	33.33
62.33	3	7.14	1	7.14	100.00
67.49	3	7.14	1	7.14	100.00
Total	42	100.00	28	200.00	50.00
			(n = 14)		

图 21.17　面板数据回归分析结果 6

图 21.18 是“top5”变量组内、组间以及整体的分布频率的结果。

. xttab top5

top5	Overall Freq.	Overall Percent	Between Freq.	Between Percent	Within Percent
.74	1	2.38	1	7.14	33.33
1.89	1	2.38	1	7.14	33.33
4.14	1	2.38	1	7.14	33.33
7.62	1	2.38	1	7.14	33.33
7.7	1	2.38	1	7.14	33.33
8.52	1	2.38	1	7.14	33.33
10.23	1	2.38	1	7.14	33.33
17.73	1	2.38	1	7.14	33.33
19.43	1	2.38	1	7.14	33.33
19.95	1	2.38	1	7.14	33.33
21.89	1	2.38	1	7.14	33.33
23.65	1	2.38	1	7.14	33.33
24.03	1	2.38	1	7.14	33.33
25.5	1	2.38	1	7.14	33.33
26.14	1	2.38	1	7.14	33.33
27.54	1	2.38	1	7.14	33.33
33.18	1	2.38	1	7.14	33.33
37.4	1	2.38	1	7.14	33.33
40.29	3	7.14	1	7.14	100.00
40.63	1	2.38	1	7.14	33.33
41.18	1	2.38	1	7.14	33.33
41.91	1	2.38	1	7.14	33.33
42.28	1	2.38	1	7.14	33.33
43.22	1	2.38	1	7.14	33.33
43.37	1	2.38	1	7.14	33.33
44.42	1	2.38	1	7.14	33.33
70	1	2.38	1	7.14	33.33
70.02	1	2.38	1	7.14	33.33
70.03	1	2.38	1	7.14	33.33
93.05	2	4.76	1	7.14	66.67
93.07	1	2.38	1	7.14	33.33
94.38	1	2.38	1	7.14	33.33
94.39	1	2.38	1	7.14	33.33
94.73	1	2.38	1	7.14	33.33
94.74	2	4.76	1	7.14	66.67
95.3	1	2.38	1	7.14	33.33
95.45	1	2.38	1	7.14	33.33
95.46	1	2.38	1	7.14	33.33
Total	42	100.00	38	271.43	36.84
			(n = 14)		

图 21.18　面板数据回归分析结果 7

图 21.19 是“stop1”变量组内、组间以及整体的分布频率的结果。

图 21.20 是对每个个体显示“roe”变量的时间序列图的结果。

. xttab stop1

stop1	Overall Freq.	Overall Percent	Between Freq.	Between Percent	Within Percent
.000841	1	2.38	1	7.14	33.33
.006724	1	2.38	1	7.14	33.33
.016129	1	2.38	1	7.14	33.33
.026244	1	2.38	1	7.14	33.33
.060025	1	2.38	1	7.14	33.33
.093636	1	2.38	1	7.14	33.33
.107584	1	2.38	1	7.14	33.33
.25	1	2.38	1	7.14	33.33
.3481	2	4.76	1	7.14	66.67
.429025	1	2.38	1	7.14	33.33
.579121	1	2.38	1	7.14	33.33
1.038361	1	2.38	1	7.14	33.33
1.1664	1	2.38	1	7.14	33.33
1.456849	1	2.38	1	7.14	33.33
1.633284	2	4.76	1	7.14	66.67
1.687401	1	2.38	1	7.14	33.33
2.582449	3	7.14	1	7.14	100.00
2.808976	1	2.38	1	7.14	33.33
3.1684	1	2.38	1	7.14	33.33
3.196944	2	4.76	1	7.14	66.67
4.1616	1	2.38	1	7.14	33.33
5.555449	2	4.76	1	7.14	66.67
7.011904	3	7.14	1	7.14	100.00
12.48209	3	7.14	1	7.14	100.00
31.97902	1	2.38	1	7.14	33.33
34.95174	1	2.38	1	7.14	33.33
38.85029	3	7.14	1	7.14	100.00
45.549	3	7.14	1	7.14	100.00
Total	42	100.00	28	200.00	50.00
			(n = 14)		

图 21.19　面板数据回归分析结果 8

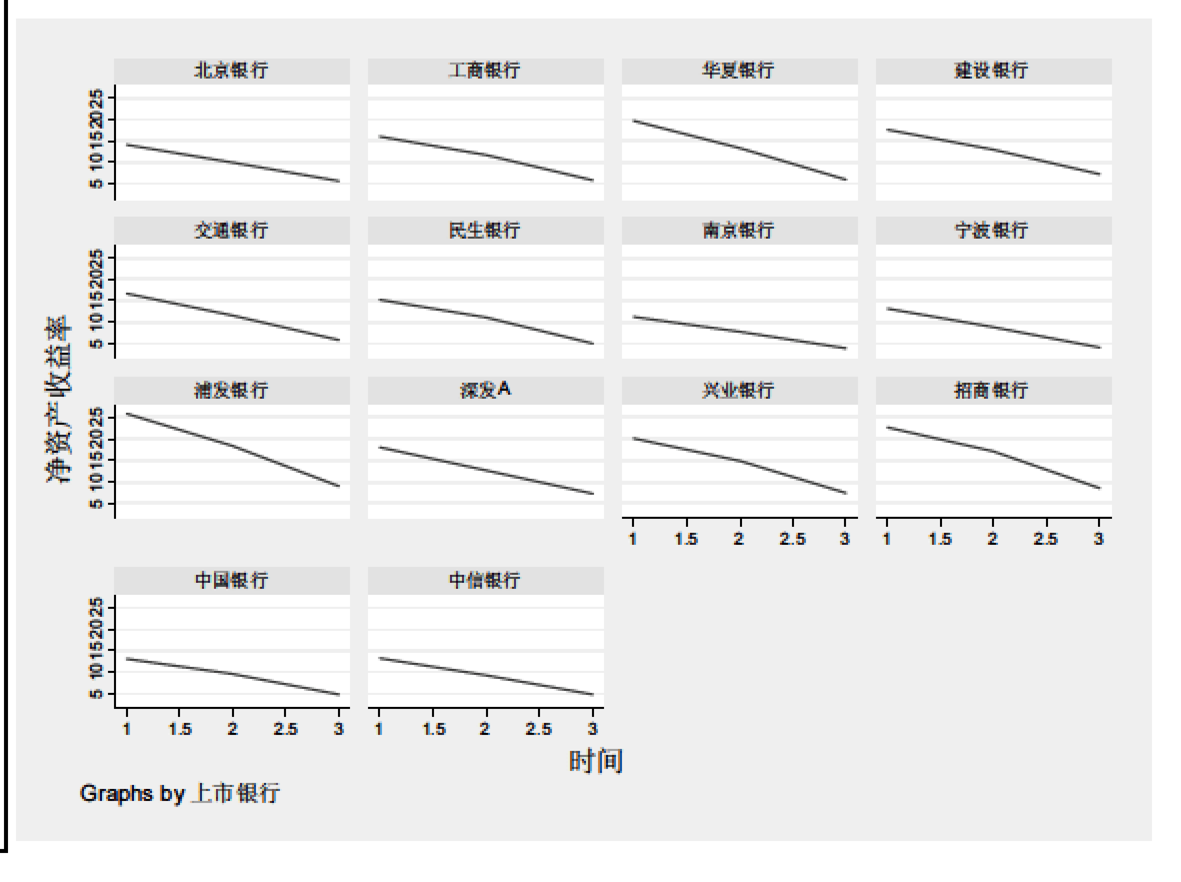


图 21.20　面板数据回归分析结果 9

从图 21.20 可以看出，不同银行的净资产收益率的时间趋势是大致相同的，都随着时间的推移而下降，但是下降的速度和平缓程度存在一定的差别。

图 21.21 是对每个个体显示“top1”变量的时间序列图的结果。

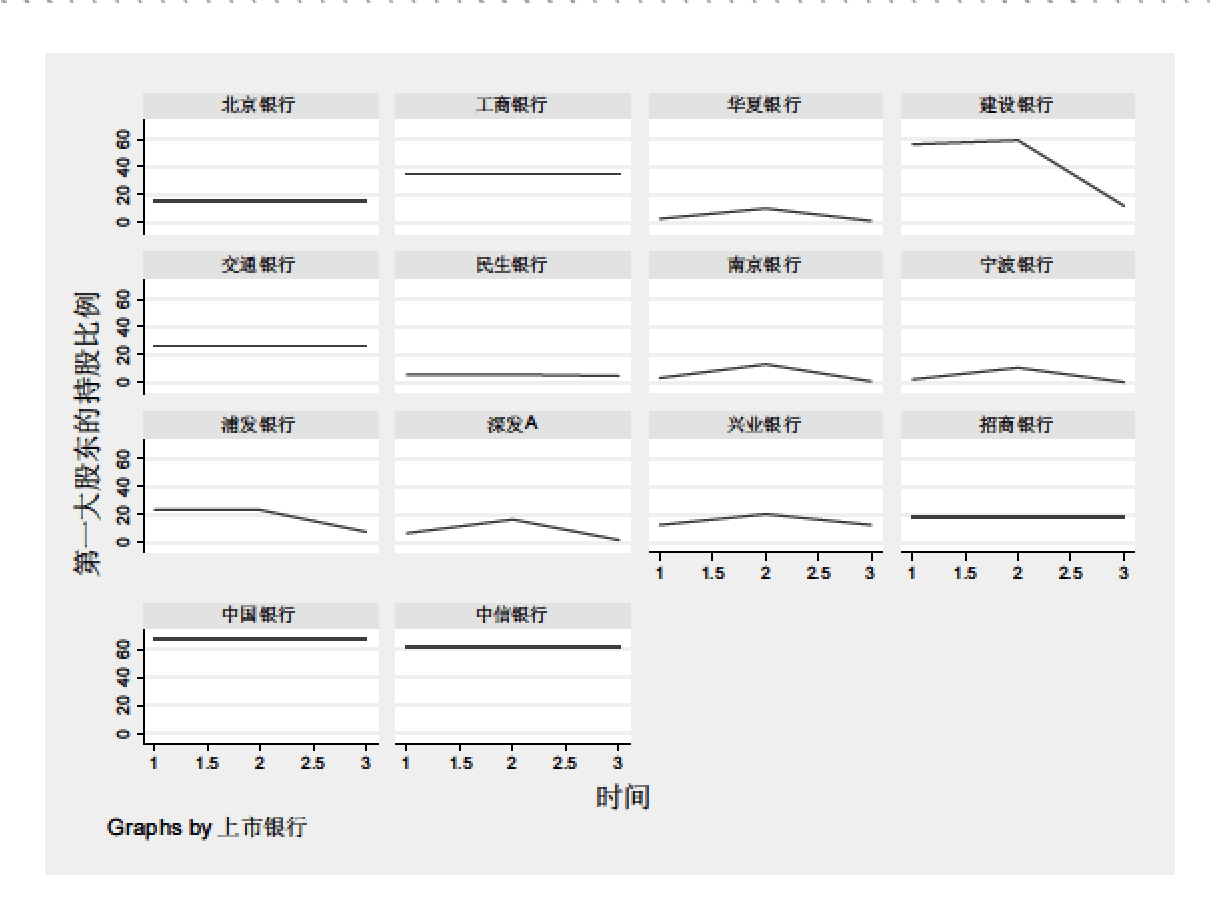


图 21.21　面板数据回归分析结果 10

从图 21.21 可以看出，不同银行的第一大股东的持股比率的时间趋势是不一致的，有的银行是持续不变的，有的是先上升后下降，有的是先不变后下降。

图 21.22 是对每个个体显示“top5”变量的时间序列图的结果。

从图 21.22 可以看出，不同银行的前五大股东的持股比率的时间趋势是不一致的，有的银行是持续不变的，有的是先上升后下降，有的是先不变后下降。

图 21.23 是对每个个体显示“stop1”变量的时间序列图的结果。

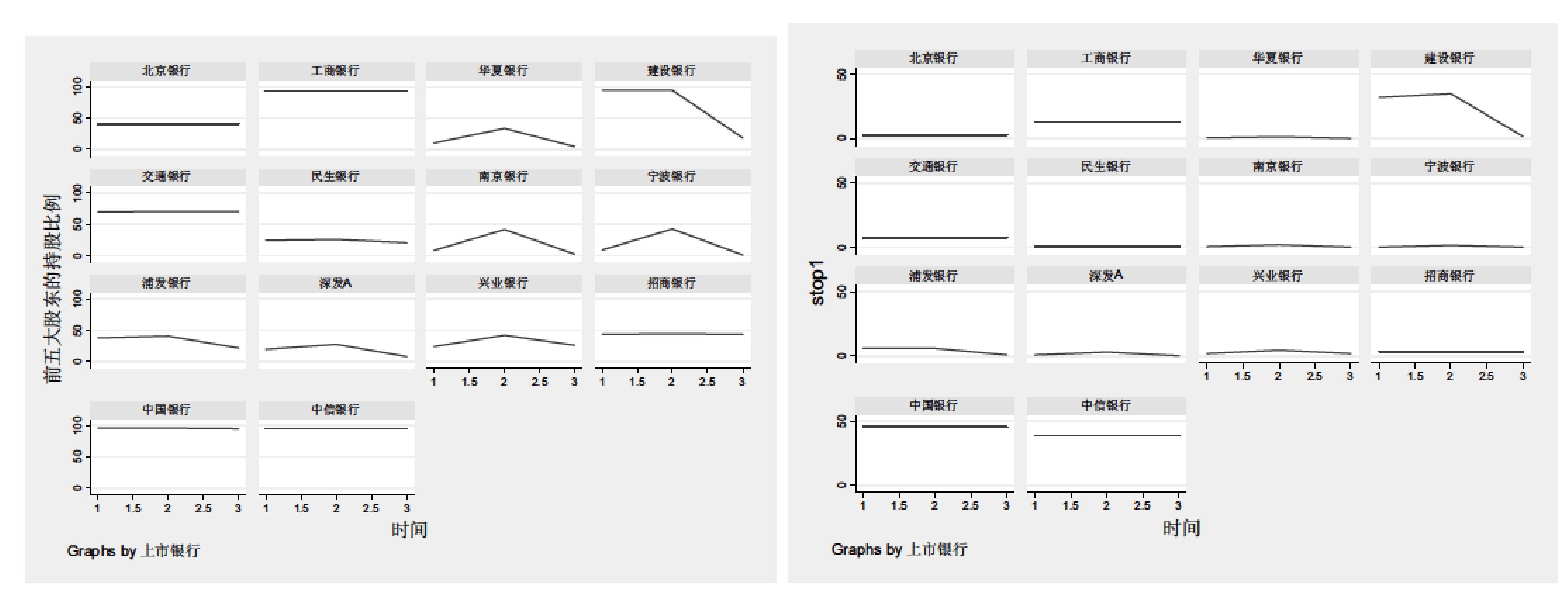


图 21.22　面板数据回归分析结果 11　　　图 21.23　面板数据回归分析结果 12

从图 21.23 可以看出，不同银行的第一大股东的持股比率的平方的时间趋势是不一致的，有的银行是持续不变的，有的是先上升后下降，有的是先不变后下降。

图 21.24 是以 roe 为因变量，以 top1、top5、stop1 为自变量，并使用以“bank”为聚类变量的聚类稳健标准差，进行固定效应回归分析的结果。

```
. xtreg roe top1 top5 stop1,fe vce(cluster  bank)

Fixed-effects (within) regression               Number of obs      =        42
Group variable: bank                            Number of groups   =        14

R-sq:  within  = 0.1698                         Obs per group: min =         3
       between = 0.0028                                        avg =       3.0
       overall = 0.0241                                        max =         3

                                                F(3,13)            =    363.38
corr(u_i, Xb)  = -0.7945                        Prob > F           =    0.0000

                                 (Std. Err. adjusted for 14 clusters in bank)

             |               Robust
         roe |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        top1 |   .9494066   .5814926     1.63   0.127    -.3068317    2.205645
        top5 |  -.1757092   .1732424    -1.01   0.329    -.5499767    .1985583
       stop1 |   -.698867   .4274474    -1.63   0.126    -1.622311     .224577
       _cons |   4.626814   1.532067     3.02   0.010     1.316985    7.936644
-------------+----------------------------------------------------------------
     sigma_u |  5.5670591
     sigma_e |  5.3815613
         rho |  .51693771   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

图 21.24　面板数据回归分析结果 13

从图 21.24 中可以看到共有 14 组，每组 3 个，共有 42 个样本参与了固定效应回归分析。模型的 F 值是 363.38，显著性 P 值为 0.0000，模型是非常显著的。模型组内 R 方是 0.1698（within = 0.1698），说明单位内解释的变化比例是 16.98%。模型组间 R 方是 0.0028（within = 0.0028），说明单位间解释的变化比例是 0.28%。模型总体 R 方是 0.0241（overall = 0.0241），说明总的解释变化比例是 2.41%。模型的解释能力不够良好。观察模型中各个变量系数的显著性 P 值，发现也都是比较显著的。此外观察图 21.24 中最后一行，rho=0.51693771，说明复合扰动项的方差也有一部分属于时间效应的变动，这一点在后面的分析中也可以得到验证。

图 21.25 是以 roe 为因变量，以 top1、top5、stop1 为自变量，进行固定效应回归分析的结果。

```
. xtreg roe top1 top5 stop1,fe

Fixed-effects (within) regression               Number of obs      =        42
Group variable: bank                            Number of groups   =        14

R-sq:  within  = 0.1698                         Obs per group: min =         3
       between = 0.0028                                        avg =       3.0
       overall = 0.0241                                        max =         3

                                                F(3,25)            =      1.70
corr(u_i, Xb)  = -0.7945                        Prob > F           =    0.1917

------------------------------------------------------------------------------
         roe |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        top1 |   .9494066   .6397925     1.48   0.150    -.3682707    2.267084
        top5 |  -.1757092   .2154388    -0.82   0.422    -.6194136    .2679953
       stop1 |   -.698867   .6238427    -1.12   0.273    -1.983695    .5859612
       _cons |   4.626814   3.563254     1.30   0.206    -2.711846    11.96547
-------------+----------------------------------------------------------------
     sigma_u |  5.5670591
     sigma_e |  5.3815613
         rho |  .51693771   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(13, 25) =     0.61             Prob > F = 0.8264
```

图 21.25　面板数据回归分析结果 14

本结果相对于使用以“bank”为聚类变量的聚类稳健标准差进行固定效应回归分析的结果在变量系数显著性上有所降低。此外，在图 21.25 的最下面一行，可以看到“F test that all u_i=0:

F(13, 25) = 0.61 Prob > F = 0.8264”，即显著接受了各个样本都没有自己的截距项的原假设，所以可以初步认为每个个体可以共用同一个截距项，也就是说固定效应模型是不一定优于普通最小二乘回归模型的。这一点也在后续的深入分析中得到了验证。

图 21.26 存储的是固定效应回归分析估计结果。选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 21.26 所示的变量_est_fe 的相关数据。

	stop5	stop10	bank	t1	t2	t3	_est_re	_est_fe
1	16.23284	23.28062	北京银行	1	0	0	1	1
2	16.23284	23.28062	北京银行	0	1	0	1	1
3	16.23284	23.28062	北京银行	0	0	1	1	1
4	86.62025	91.83389	工商银行	1	0	0	1	1
5	86.58303	91.9681	工商银行	0	1	0	1	1
6	86.58303	91.9681	工商银行	0	0	1	1	1
7	1.046529	1.954404	华夏银行	1	0	0	1	1
8	11.00912	22.06181	华夏银行	0	1	0	1	1
9	.171396	.588289	华夏银行	0	0	1	1	1
10	89.07584	92.62137	建设银行	1	0	0	1	1
11	89.09472	92.65988	建设银行	0	1	0	1	1
12	3.143529	3.265249	建设银行	0	0	1	1	1
13	49	53.47997	交通银行	1	0	0	1	1
14	49.04201	53.52386	交通银行	0	1	0	1	1
15	49.028	53.50923	交通银行	0	0	1	1	1
16	5.593225	14.49325	民生银行	1	0	0	1	1
17	6.5025	18.74024	民生银行	0	1	0	1	1
18	3.980025	10.26562	民生银行	0	0	1	1	1
19	.5929	1.234321	南京银行	1	0	0	1	1
20	16.95792	21.27054	南京银行	0	1	0	1	1
21	.035721	.0676	南京银行	0	0	1	1	1
22	.7259041	.998001	宁波银行	1	0	0	1	1
23	17.87598	38.52685	宁波银行	0	1	0	1	1
24	.005476	.010609	宁波银行	0	0	1	1	1
25	13.9876	19.0969	浦发银行	1	0	0	1	1
26	16.50797	21.24288	浦发银行	0	1	0	1	1
27	4.791721	8.04857	浦发银行	0	0	1	1	1
28	3.775249	7.263025	深发A	1	0	0	1	1
29	7.584517	11.58041	深发A	0	1	0	1	1
30	.580644	1.920996	深发A	0	0	1	1	1

图 21.26 面板数据回归分析结果 15

图 21.27 是构建最小二乘虚拟变量模型来分析固定效应模型是否优于最小二乘回归分析的分析结果。

```
. xi:xtreg roe top1 top5 stop1 i.bank,vce(cluster bank)
i.bank            _Ibank_1-14         (naturally coded; _Ibank_1 omitted)

Random-effects GLS regression                   Number of obs      =        42
Group variable: bank                            Number of groups   =        14

R-sq:  within  = 0.1698                         Obs per group: min =         3
       between = 1.0000                                        avg =       3.0
       overall = 0.3960                                        max =         3

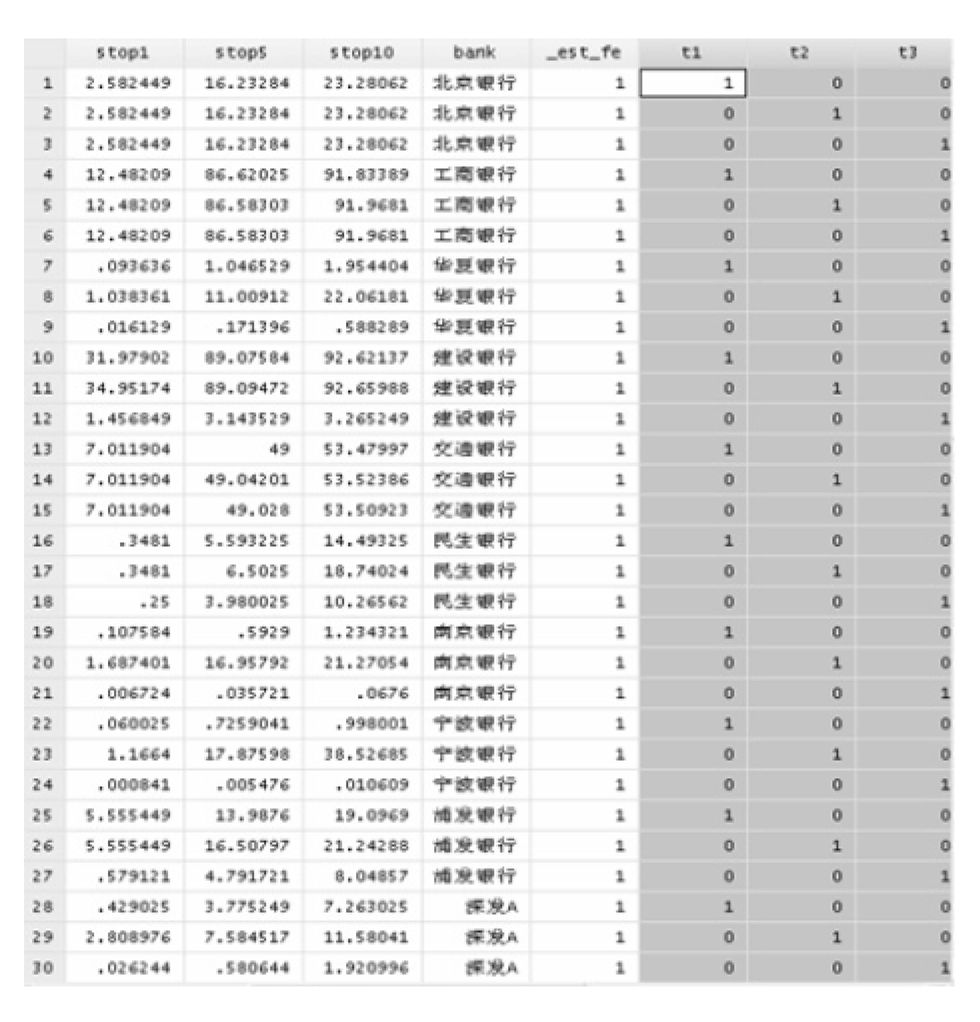
                                                Wald chi2(3)       =         .
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =         .

                                  (Std. Err. adjusted for 14 clusters in bank)
------------------------------------------------------------------------------
             |               Robust
         roe |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        top1 |   .9494066   .7169122     1.32   0.185    -.4557155    2.354529
        top5 |  -.1757092   .2135876    -0.82   0.411    -.5943332    .2429148
       stop1 |   -.698867   .5269926    -1.33   0.185    -1.731753    .3340195
    _Ibank_2 |   -.805453   2.696805    -0.30   0.765    -6.091094    4.480188
    _Ibank_3 |   7.933511   1.866924     4.25   0.000     4.274406    11.59262
    _Ibank_4 |  -3.336949   2.501829    -1.33   0.182    -8.240443    1.566546
    _Ibank_5 |  -.1411415   1.303835    -0.11   0.914    -2.696612    2.414329
    _Ibank_6 |   5.860433    2.71435     2.16   0.031     .5404051    11.18046
    _Ibank_7 |    2.12437   1.605233     1.32   0.186    -1.021829    5.270568
    _Ibank_8 |    4.19578   2.357546     1.78   0.075    -.4249251    8.816485
    _Ibank_9 |   5.478392   2.373487     2.31   0.021      .826444    10.13034
   _Ibank_10 |   5.177696   .6416579     8.07   0.000     3.920069    6.435322
   _Ibank_11 |   3.221478   1.573751     2.05   0.041     .1369832    6.305972
   _Ibank_12 |   5.533589   .2591392    21.35   0.000     5.025686    6.041493
   _Ibank_13 |  -9.806663   3.136047    -3.13   0.002     -15.9532   -3.660124
   _Ibank_14 |  -9.759707   2.901101    -3.36   0.001    -15.44576   -4.073654
       _cons |   3.507148   2.036352     1.72   0.085    -.4840296    7.498325
-------------+----------------------------------------------------------------
     sigma_u |          0
     sigma_e |  5.3815613
         rho |          0   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

图 21.27 面板数据回归分析结果 16

从图 21.27 中可以看出，大多数个体虚拟变量的显著性 P 值都是大于 0.05 的，所以可以在一定程度上认为可以接受“所有个体的虚拟变量皆为 0”的原假设，也就是说固定效应模型不一定是优于普通最小二乘回归模型的。

图 21.28 是创建年度变量的多个虚拟变量的结果。选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 21.28 所示的变量 t1~t3 的相关数据。



	stop1	stop5	stop10	bank	_est_fe	t1	t2	t3
1	2.582449	16.23284	23.28062	北京银行	1	1	0	0
2	2.582449	16.23284	23.28062	北京银行	1	0	1	0
3	2.582449	16.23284	23.28062	北京银行	1	0	0	1
4	12.48209	86.62025	91.83389	工商银行	1	1	0	0
5	12.48209	86.58303	91.9681	工商银行	1	0	1	0
6	12.48209	86.58303	91.9681	工商银行	1	0	0	1
7	.093636	1.046529	1.954404	华夏银行	1	1	0	0
8	1.038361	11.00912	22.06181	华夏银行	1	0	1	0
9	.016129	.171396	.588289	华夏银行	1	0	0	1
10	31.97902	89.07584	92.62137	建设银行	1	1	0	0
11	34.95174	89.09472	92.65988	建设银行	1	0	1	0
12	1.456849	3.143529	3.265249	建设银行	1	0	0	1
13	7.011904	49	53.47997	交通银行	1	1	0	0
14	7.011904	49.04201	53.52386	交通银行	1	0	1	0
15	7.011904	49.028	53.50923	交通银行	1	0	0	1
16	.3481	5.593225	14.49325	民生银行	1	1	0	0
17	.3481	6.5025	18.74024	民生银行	1	0	1	0
18	.25	3.980025	10.26562	民生银行	1	0	0	1
19	.107584	.5929	1.234321	南京银行	1	1	0	0
20	1.687401	16.95792	21.27054	南京银行	1	0	1	0
21	.006724	.035721	.0676	南京银行	1	0	0	1
22	.060025	.7259041	.998001	宁波银行	1	1	0	0
23	1.1664	17.87598	38.52685	宁波银行	1	0	1	0
24	.000841	.005476	.010609	宁波银行	1	0	0	1
25	5.555449	13.9876	19.0969	浦发银行	1	1	0	0
26	5.555449	16.50797	21.24288	浦发银行	1	0	1	0
27	.579121	4.791721	8.04857	浦发银行	1	0	0	1
28	.429025	3.775249	7.263025	深发A	1	1	0	0
29	2.808976	7.584517	11.58041	深发A	1	0	1	0
30	.026244	.580644	1.920996	深发A	1	0	0	1

图 21.28　面板数据回归分析结果 17

图 21.29 是构建双向固定效应模型的分析结果。

```
. xtreg roe top1 top5 stop1 t2-t3,fe vce(cluster bank)

Fixed-effects (within) regression               Number of obs      =        42
Group variable: bank                            Number of groups   =        14

R-sq:  within  = 0.9510                         Obs per group: min =         3
       between = 0.0759                                        avg =       3.0
       overall = 0.7115                                        max =         3

                                                F(5,13)            =    219.32
corr(u_i, Xb)  = -0.0135                        Prob > F           =    0.0000

                                 (Std. Err. adjusted for 14 clusters in bank)
------------------------------------------------------------------------------
             |               Robust
         roe |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        top1 |   .3322207   .2871586     1.16   0.268    -.2881478    .9525891
        top5 |  -.0938745   .0713693    -1.32   0.211    -.2480585    .0603094
       stop1 |  -.2444363   .2394989    -1.02   0.326    -.7618421    .2729695
          t2 |  -4.864719   .4090549   -11.89   0.000    -5.748428   -3.981009
          t3 |  -10.56468   .7177505   -14.72   0.000    -12.11528   -9.014072
       _cons |   15.88513   1.244676    12.76   0.000     13.19617    18.57409
-------------+----------------------------------------------------------------
     sigma_u |  2.7877008
     sigma_e |  1.3631971
         rho |  .80702108   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

图 21.29　面板数据回归分析结果 18

从图 21.29 中可以看出，全部虚拟变量的显著性 P 值都是远小于 0.05 的，所以可以初步认为模型中应该包含时间效应。值得说明的是，在构建双向固定效应模型时并没有把 t1 列入进去，这是因为 t1 被视为基期，也就是模型中的常数项。

包含时间效应项的模型的回归方程是：

ROE=0.3322207 *top1 -0.2444363 *stop1-0.0938745*top5-4.864719*t2-10.56468*t3+15.88513

可以看出 stop1 前面的系数为负，说明中国上市银行的 ROE 与第一大股东持股量之间存在着倒“U”型关系。t2、t3 前面的系数显著为负，而且 t3 的负程度更大，说明随着时间的推移，净资产收益率是不断下降的。

图 21.30 是在上步回归的基础上，通过测试各虚拟变量的系数联合显著性来检验是否应该在模型中纳入时间效应的检验结果。

从图 21.30 中可以看出，各变量系数的联合显著性是非常差的，即强烈拒绝了没有时间效应的初始假设，所以，我们进一步验证了模型中应该包含时间效应项的结论。

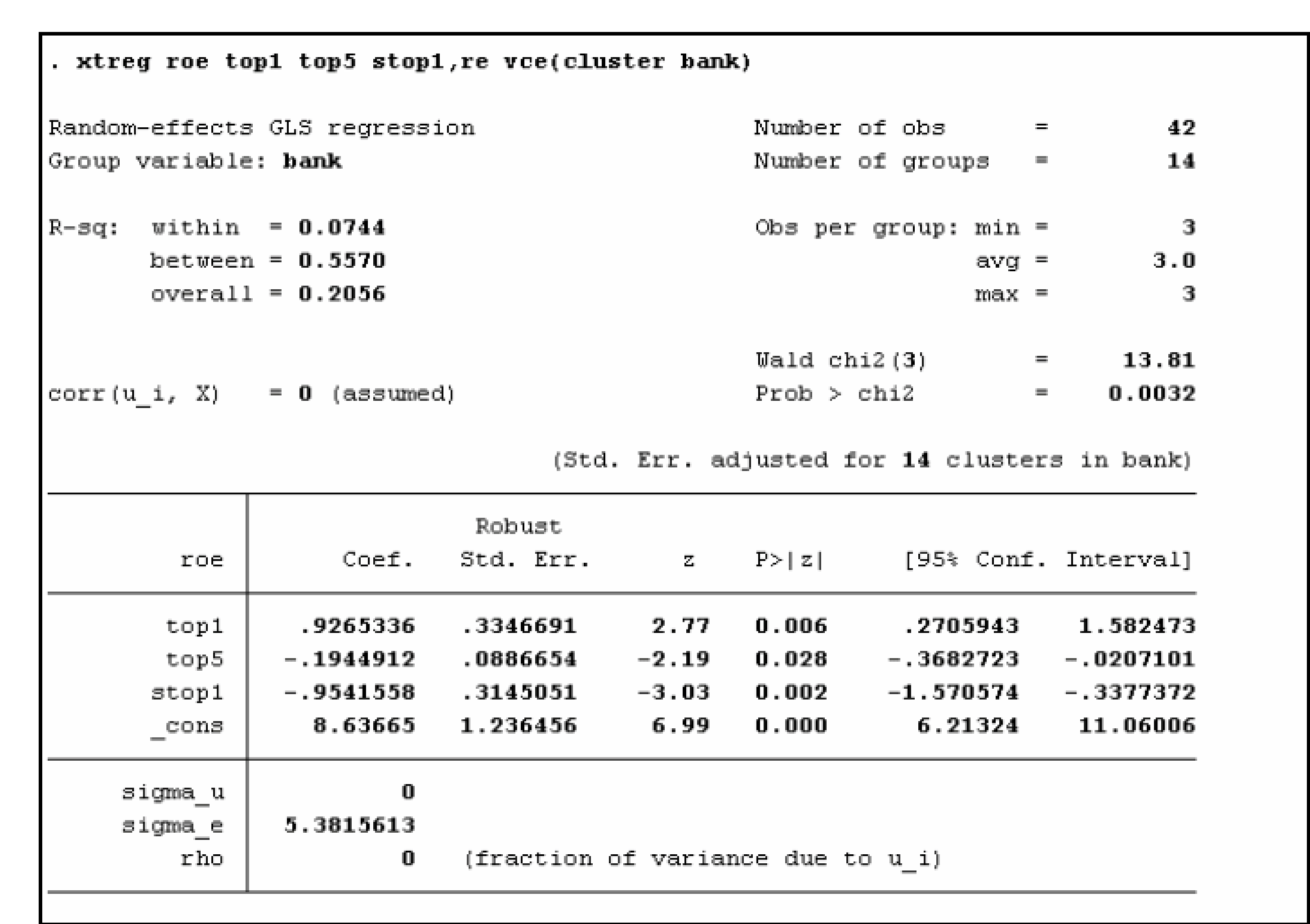
图 21.31 是以 roe 为因变量，以 top1、top5、stop1 为自变量，并使用以“bank”为聚类变量的聚类稳健标准差进行随机效应回归分析的结果。

```
. test  t2 t3

 ( 1)  t2 = 0
 ( 2)  t3 = 0

       F(  2,    13) =  116.05
            Prob > F =   0.0000
```

图 21.30　面板数据回归分析结果 19



```
. xtreg roe top1 top5 stop1,re vce(cluster bank)

Random-effects GLS regression                   Number of obs      =        42
Group variable: bank                            Number of groups   =        14

R-sq:  within  = 0.0744                         Obs per group: min =         3
       between = 0.5570                                        avg =       3.0
       overall = 0.2056                                        max =         3

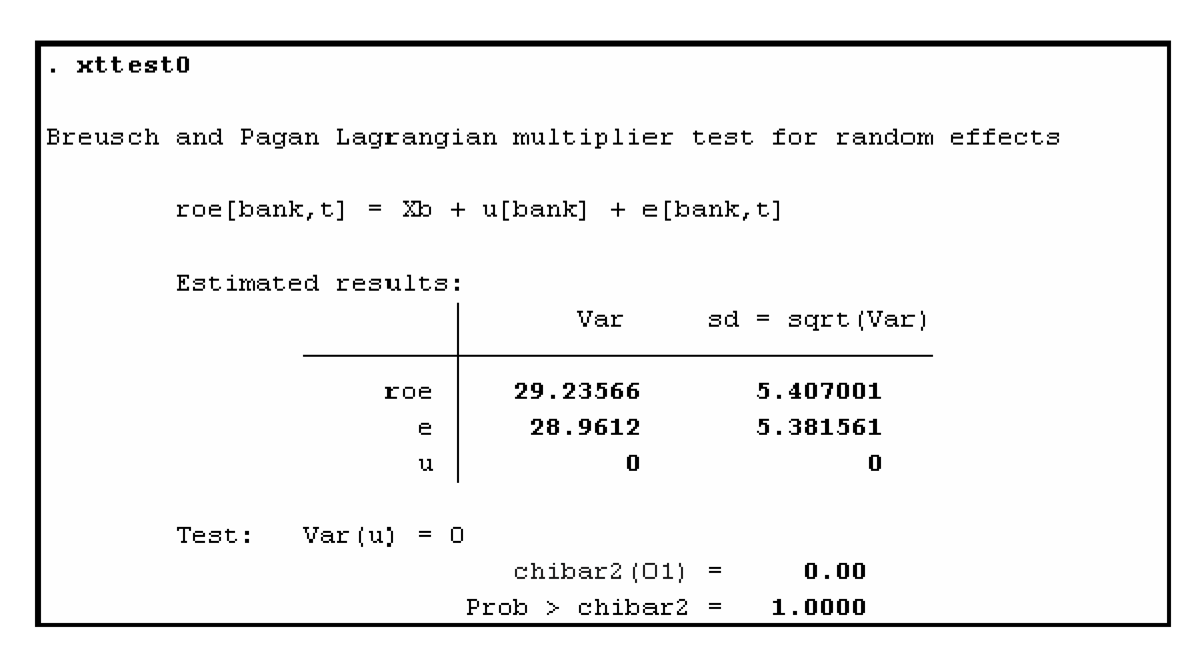
                                                Wald chi2(3)       =     13.81
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0032

                                 (Std. Err. adjusted for 14 clusters in bank)
------------------------------------------------------------------------------
             |               Robust
         roe |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        top1 |   .9265336   .3346691     2.77   0.006     .2705943    1.582473
        top5 |  -.1944912   .0886654    -2.19   0.028    -.3682723   -.0207101
       stop1 |  -.9541558   .3145051    -3.03   0.002    -1.570574   -.3377372
       _cons |    8.63665   1.236456     6.99   0.000      6.21324    11.06006
-------------+----------------------------------------------------------------
     sigma_u |          0
     sigma_e |  5.3815613
         rho |          0   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

图 21.31　面板数据回归分析结果 20

从图 21.31 可以看出，随机效应回归分析的结果相比固定效应回归分析在变量的显著性水平上得到了大幅度的提高，变量系数显著性变得非常好。

图 21.32 是在上步回归的基础上，进行假设检验来判断随机效应模型是否优于最小二乘回归模型的结果。



```
. xttest0

Breusch and Pagan Lagrangian multiplier test for random effects

        roe[bank,t] = Xb + u[bank] + e[bank,t]

        Estimated results:
                         |       Var     sd = sqrt(Var)
                ---------+-----------------------------
                     roe |   29.23566       5.407001
                       e |    28.9612       5.381561
                       u |          0              0

        Test:   Var(u) = 0
                             chibar2(01) =     0.00
                          Prob > chibar2 =   1.0000
```

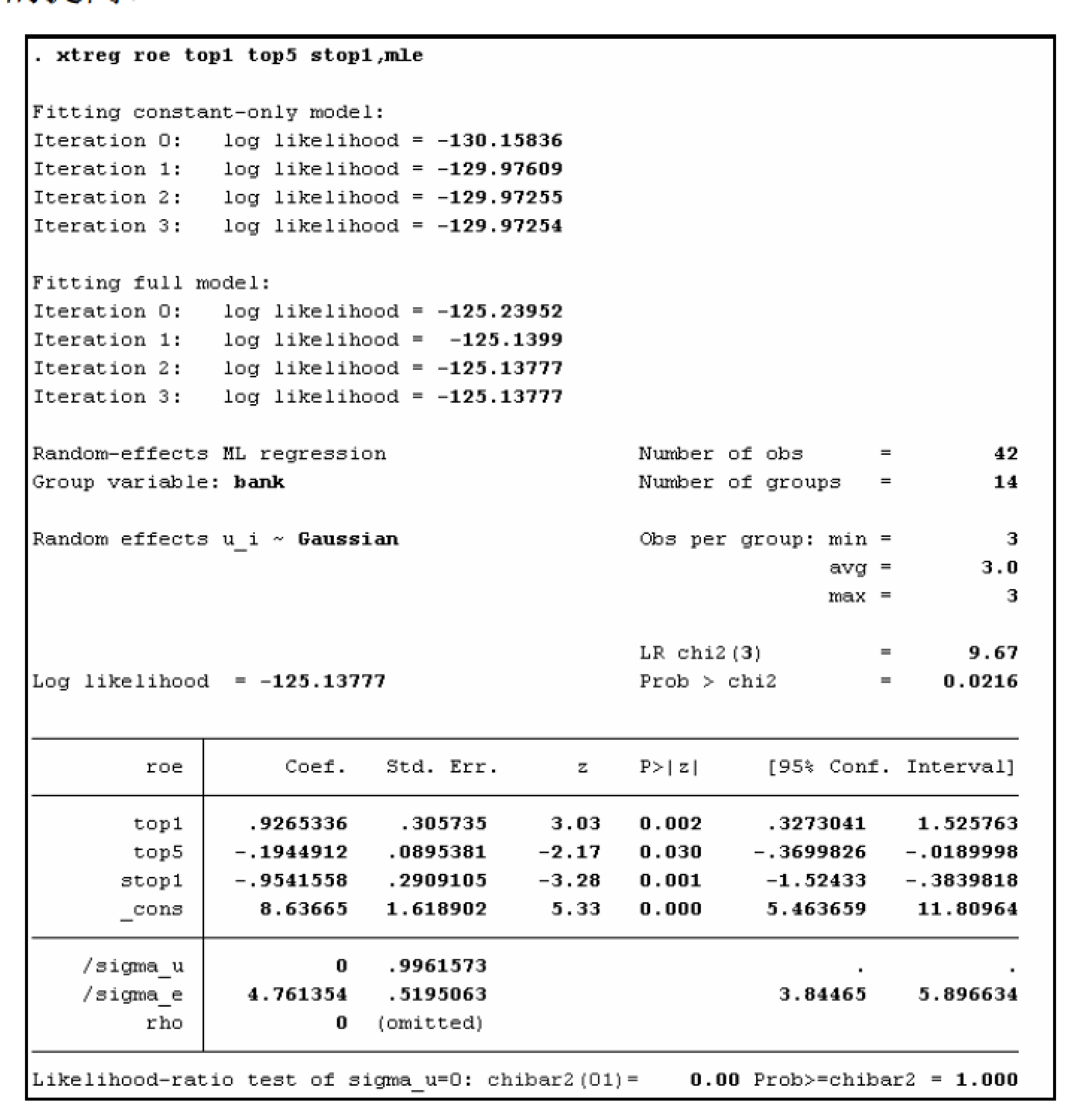
图 21.32　面板数据回归分析结果 21

从图 21.32 可以看出，假设检验非常显著地接受了不存在个体随机效应的原假设，也就是

说，随机效应模型并不优于普通最小二乘回归分析模型。

图 21.33 是以 roe 为因变量，以 top1、top5、stop1 为自变量，并使用最大似然估计方法，进行随机效应回归分析的结果。

从图 21.33 可以看出，使用最大似然估计方法的随机效应回归分析的结果与使用以“bank”为聚类变量的聚类稳健标准差的随机效应回归分析的结果大同小异，只是部分变量的显著性水平得到了进一步的提高。



```
. xtreg roe top1 top5 stop1,mle

Fitting constant-only model:
Iteration 0:   log likelihood = -130.15836
Iteration 1:   log likelihood = -129.97609
Iteration 2:   log likelihood = -129.97255
Iteration 3:   log likelihood = -129.97254

Fitting full model:
Iteration 0:   log likelihood = -125.23952
Iteration 1:   log likelihood =  -125.1399
Iteration 2:   log likelihood = -125.13777
Iteration 3:   log likelihood = -125.13777

Random-effects ML regression                    Number of obs      =        42
Group variable: bank                            Number of groups   =        14

Random effects u_i ~ Gaussian                   Obs per group: min =         3
                                                               avg =       3.0
                                                               max =         3

                                                LR chi2(3)         =      9.67
Log likelihood  = -125.13777                    Prob > chi2        =    0.0216

------------------------------------------------------------------------------
         roe |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        top1 |   .9265336    .305735     3.03   0.002     .3273041    1.525763
        top5 |  -.1944912   .0895381    -2.17   0.030    -.3699826   -.0189998
       stop1 |  -.9541558   .2909105    -3.28   0.001     -1.52433   -.3839818
       _cons |    8.63665   1.618902     5.33   0.000     5.463659    11.80964
-------------+----------------------------------------------------------------
    /sigma_u |          0   .9961573                             .           .
    /sigma_e |   4.761354   .5195063                       3.84465    5.896634
         rho |          0  (omitted)
------------------------------------------------------------------------------
Likelihood-ratio test of sigma_u=0: chibar2(01)=     0.00 Prob>=chibar2 = 1.000
```

图 21.33　面板数据回归分析结果 22

图 21.34 是以 roe 为因变量，以 top1、top5、stop1 为自变量，并使用组间估计量，进行组间估计量回归分析的结果。

```
. xtreg roe top1 top5 stop1,be

Between regression (regression on group means)  Number of obs      =        42
Group variable: bank                            Number of groups   =        14

R-sq:  within  = 0.0473                         Obs per group: min =         3
       between = 0.5969                                        avg =       3.0
       overall = 0.1936                                        max =         3

                                                F(3,10)            =      4.94
sd(u_i + avg(e_i.))=  2.094773                  Prob > F           =    0.0235

------------------------------------------------------------------------------
         roe |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        top1 |   .9784151   .2792529     3.50   0.006     .3562008    1.600629
        top5 |  -.2311831   .0800543    -2.89   0.016    -.4095552   -.0528109
       stop1 |  -.9888918   .2651622    -3.73   0.004     -1.57971   -.3980737
       _cons |   9.495548   1.375315     6.90   0.000     6.431154    12.55994
------------------------------------------------------------------------------
```

图 21.34　面板数据回归分析结果 23

从图 21.34 可以看出，使用组间估计量进行回归分析的结果与固定效应模型、随机效应模型在模型的解释能力以及变量系数的显著性上都大同小异。

图 21.35 是以 roe 为因变量，以 top1、top5、stop1 为自变量，进行随机效应回归分析的结果。

对该回归分析结果的详细解读在前面也多次讲述，此处不再重复讲解。

```
. xtreg roe top1 top5 stop1,re

Random-effects GLS regression                   Number of obs      =        42
Group variable: bank                            Number of groups   =        14

R-sq:  within  = 0.0744                         Obs per group: min =         3
       between = 0.5570                                        avg =       3.0
       overall = 0.2056                                        max =         3

                                                Wald chi2(3)       =      9.84
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0200

------------------------------------------------------------------------------
         roe |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        top1 |   .9265336   .3214238     2.88   0.004     .2965546    1.556513
        top5 |  -.1944912   .0941327    -2.07   0.039     -.378988   -.0099945
       stop1 |  -.9541558   .3058385    -3.12   0.002    -1.553588   -.3547233
       _cons |    8.63665   1.701976     5.07   0.000     5.300837    11.97246
-------------+----------------------------------------------------------------
     sigma_u |          0
     sigma_e |  5.3015613
         rho |          0   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

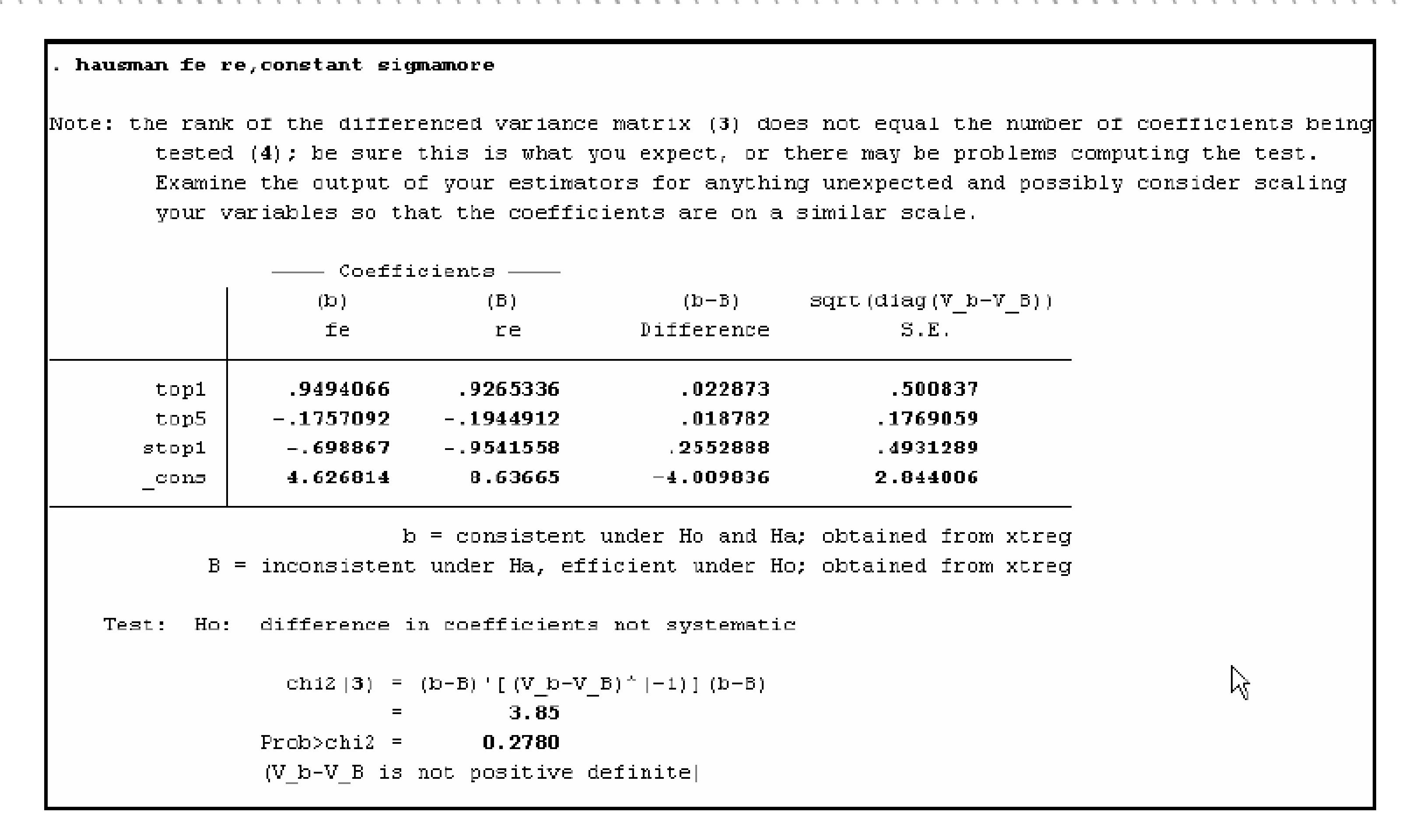
图 21.35　面板数据回归分析结果 24

图 21.36 存储的是随机效应回归分析估计结果。选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 21.36 所示的变量_est_re 的相关数据。

	stop5	stop10	bank	_est_fe	t1	t2	t3	_est_re
1	16.23284	23.28062	北京银行	1	1	0	0	1
2	16.23284	23.28062	北京银行	1	0	1	0	1
3	16.23284	23.28062	北京银行	1	0	0	1	1
4	86.62025	91.83389	工商银行	1	1	0	0	1
5	86.58303	91.9681	工商银行	1	0	1	0	1
6	86.58303	91.9681	工商银行	1	0	0	1	1
7	1.046529	1.954404	华夏银行	1	1	0	0	1
8	11.00912	22.06181	华夏银行	1	0	1	0	1
9	.171396	.588289	华夏银行	1	0	0	1	1
10	89.07584	92.62137	建设银行	1	1	0	0	1
11	89.09472	92.65988	建设银行	1	0	1	0	1
12	3.143529	3.265249	建设银行	1	0	0	1	1
13	49	53.47997	交通银行	1	1	0	0	1
14	49.04201	53.52386	交通银行	1	0	1	0	1
15	49.028	53.50923	交通银行	1	0	0	1	1
16	5.593225	14.49325	民生银行	1	1	0	0	1
17	6.5025	18.74024	民生银行	1	0	1	0	1
18	3.980025	10.26562	民生银行	1	0	0	1	1
19	.5929	1.234321	南京银行	1	1	0	0	1
20	16.95792	21.27054	南京银行	1	0	1	0	1
21	.035721	.0676	南京银行	1	0	0	1	1
22	.7259041	.998001	宁波银行	1	1	0	0	1
23	17.87598	38.52685	宁波银行	1	0	1	0	1
24	.005476	.010609	宁波银行	1	0	0	1	1
25	13.9876	19.0969	浦发银行	1	1	0	0	1
26	16.50797	21.24288	浦发银行	1	0	1	0	1
27	4.791721	8.04857	浦发银行	1	0	0	1	1
28	3.775249	7.263025	深发A	1	1	0	0	1
29	7.584517	11.58041	深发A	1	0	1	0	1
30	.580644	1.920996	深发A	1	0	0	1	1

图 21.36　面板数据回归分析结果 25

图 21.37 是进行豪斯曼检验的结果。



```
. hausman fe re,constant sigmamore

Note: the rank of the differenced variance matrix (3) does not equal the number of coefficients being
      tested (4); be sure this is what you expect, or there may be problems computing the test.
      Examine the output of your estimators for anything unexpected and possibly consider scaling
      your variables so that the coefficients are on a similar scale.

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |      fe           re          Difference          S.E.
-------------+----------------------------------------------------------------
        top1 |    .9494066     .9265336        .022873         .500837
        top5 |   -.1757092    -.1944912        .018782         .1769059
       stop1 |    -.698867    -.9541558        .2552888        .4931289
       _cons |    4.626814      8.63665      -4.009836        2.844006
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from xtreg
            B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(3) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =        3.85
                Prob>chi2 =      0.2780
                (V_b-V_B is not positive definite)
```

图 21.37　面板数据回归分析结果 26

豪斯曼检验的原假设是使用随机效应模型。图 21.37 显示的显著性 P 值（Prob>chi2 =0.2780）远远大于 5%，所以我们接受初始假设，认为使用随机效应模型是更为合理的。

综上所述，我们应该构建随机效应模型或者使用普通最小二乘回归分析方法来描述变量之间的回归关系。

21.4　研究结论

从前面的分析中可以看出，不论是随机效应模型还是普通最小二乘回归模型，top1 的系数都是大于 0 的值，并且 stop1 的系数都是小于 0 的值，所以可以得出最后的结论：上市银行的净资产收益率和股权集中度之间是一种倒“U”型关系。

产生上述结果的原因可以从以下两方面来解释。

- 对倒“U”型上升阶段的解释：上市银行的主要特征是所有权和经营权分离，因此必然会产生委托代理问题。由于存在信息不对称和代理人的道德风险等问题，股东作为委托人必须对代理人进行有效的监督管理和激励约束才能促进银行绩效的提高，但如果上市银行的股权过分分散，必然使股东们“搭便车”的心态严重，不愿对上市银行治理进行改进而影响银行绩效。如果存在持股比例较大的股东，一方面，他有能力获取公司发展的最新信息，信息不对称相对不严重，对管理层的监督成本相对较低，使他容易解决好委托代理问题；另一方面，由于持股比例较大，股权流动相对较难，控股股东对公司经营和长期发展往往比较关心，有动力将公司经营好，选拔优秀管理人才，对管理层进行较有效的激励约束，甚至直接向上市公司注入优质资产进行支持。这些都有助于上市公司绩效的提高。
- 对倒“U”型下降阶段的解释：一方面，在十几家上市银行中，第一大股东持股量很多都是国家控股的，而国有银行还没有完全实现机制的转换，没有真正地把国有资产置于投资者的监督之下，国有银行的低效率残余还很浓厚；另一方面，由于我国法制的不健全，我国证券市场上绝对控股的大股东的存在会导致大股东控制上市银行、操

纵上市银行利润、占用上市银行资产等一系列现象。股权制衡能有效抑制大股东的恶性关联交易行为，提高银行的绩效。

21.5 本章习题

饮料行业的人士普遍认为，成功经营饮料公司最关键的环节在于销售，所以销售策略的思考与选择问题历来是市场专家研究的焦点。其间一个非常重要的问题是：饮料公司的利润与其销售集中度之间是否存在一定的相关性？某调研者选取了10家饮料公司在2008—2010年的有关数据作为观测样本进行研究，如表21.2所示。请读者帮助该调研者构建恰当模型描述饮料公司的利润与其销售集中度之间的合理关系。

表21.2 10家饮料公司的销售数据（2008—2010年）

饮料公司	第一大销售商的销售量/万瓶	前五大销售商的销售量/万瓶	前十大销售商的销售量/万瓶	利润/万元	时间（1代表2008年，2代表2009年，3代表2010年）
A	6.55	19.43	26.95	18.05	1
A	16.76	27.54	34.03	12.65	2
A	1.62	7.62	13.86	7.15	3
B	2.45	8.52	9.99	13.1	1
B	10.8	42.28	62.07	8.89	2
B	0.29	0.74	1.03	4.06	3
…	…	…	…	…	…
I	67.49	95.45	97.25	13.03	1
I	67.49	95.46	97.26	9.65	2
I	67.49	95.3	97.24	4.86	3
J	62.33	94.73	95.33	13.19	1
J	62.33	94.74	95.34	9.33	2
J	62.33	94.74	95.34	4.86	3

第 22 章　Stata 在农业中的应用

农业是国民经济的重要组成部分，以生产和加工农产品为主。通常情况下，农业又被更加详细地划分为种植业、水产业、渔业、林业、畜牧业、副业等。作为第一产业，农业对于整体国民经济起着无可替代的基本作用和保障作用。而专家学者们关于农业的研究也是非常多的，很多情况下会进行定量分析以获得更加有说服力的结论，其间必然涉及对大量数据的专业统计分析。Stata 作为一种优秀的计量统计分析软件，深受农业研究者的喜爱，是他们最常使用的软件之一。下面就以实例的方式来介绍一下 Stata 在农业中的应用。

22.1　研究背景

根据《中华人民共和国年鉴 2012》提供的数据（表 22.1）可以发现，无论是农、林、牧、渔业总产值还是农业、林业、牧业、渔业的分项产值都呈现出持续快速增长趋势。

表 22.1　我国历年农、林、牧、渔业总产值及分项产值数据（单位：亿元）

年份	农、林、牧、渔业总产值	农业	林业	牧业	渔业
1978	1397.0	1117.5	48.1	209.3	22.1
1980	1922.6	1454.1	81.4	354.2	32.9
1985	3619.5	2506.4	188.7	798.3	126.1
1990	7662.1	4954.3	330.3	1967.0	410.6
1991	8157.0	5146.4	367.9	2159.2	483.5
1992	9084.7	5588.0	422.6	2460.5	613.5
1993	10 995.5	6605.1	494.0	3014.4	882.0
1994	15 750.5	9169.2	611.1	4672.0	1298.2
1995	20 340.9	11 884.6	709.9	6045.0	1701.3
1996	22 353.7	13 539.8	778.0	6015.5	2020.4
1997	23 788.4	13 852.5	817.8	6835.4	2282.7
1998	24 541.9	14 241.9	851.3	7025.8	2422.9
1999	24 519.1	14 106.2	886.3	6997.6	2529.0
2000	24 915.8	13 873.6	936.5	7393.1	2712.6
2001	26 179.6	14 462.8	938.8	7963.1	2815.0
2002	27 390.8	14 931.5	1033.5	8454.6	2971.1
2003	29 691.8	14 870.1	1239.9	9538.8	3137.6
2004	36 239.0	18 138.4	1327.1	12 173.8	3605.6
2005	39 450.9	19 613.4	1425.5	13 310.8	4016.1
2006	40 810.8	21 522.3	1610.8	12 083.9	3970.5
2007	48 893.0	24 658.1	1861.6	16 124.9	4457.5
2008	58 002.2	28 044.2	2152.9	20 583.6	5203.4

（续表）

年份	农、林、牧、渔业总产值	农业	林业	牧业	渔业
2009	60 361.0	30 777.5	2193.0	19 468.4	5626.4
2010	69 319.8	36 941.1	2595.5	20 825.7	6422.4
2011	81 303.9	41 988.6	3120.7	25 770.7	7568.0

在这种大背景下对我国目前的农业进行研究，不论是对于促进我国农业又好又快地发展，还是对于充分发挥农业对于发展国民经济和改善居民生活的作用，都有着极为重要的意义。

22.2　研究方法

按照我国目前官方统计口径，农产品产量主要体现在“粮食产量”“棉花产量”“油料产量”“麻类产量”“甘蔗产量”“甜菜产量”“烟叶产量”“茶叶产量”“水果产量”等，其中粮食产量又体现在“稻谷”“小麦”“玉米”“豆类”“薯类”等作物的产量，水果产量又体现在“苹果”“柑桔”“梨”“葡萄”“香蕉”等作物的产量，油料作物又体现在“花生”“油菜籽”“芝麻”等作物的产量，所以我们在进行分析研究的时候，考虑的关于农产品的变量也与这些叙述相吻合。

本例采用的数据为我国各省市 2011 年农产品的相关数据，包括“农业总产值”“粮食产量”“棉花产量”“油料产量”“麻类产量”“甘蔗产量”“甜菜产量”“烟叶产量”“茶叶产量”“水果产量”“谷物”“稻谷”“小麦”“玉米”“豆类”“薯类”“花生”“油菜籽”“芝麻”“黄红麻”“烤烟”“苹果”“柑桔”“梨”“葡萄”“香蕉”“谷物单位面积产量”“棉花单位面积产量”“花生单位面积产量”“油菜籽单位面积产量”“芝麻单位面积产量”“黄红麻单位面积产量”“甘蔗单位面积产量”“烤烟单位面积产量”“受灾面积（千公顷）”“成灾面积（千公顷）”“甜菜单位面积产量”等。数据都摘编自《中国统计年鉴 2012》。

采用的数据分析方法主要有描述性分析、相关分析、回归分析、因子分析、聚类分析等。

基本思路是：首先使用描述性分析来描述各个变量之间的基本特征，为后面的分析做好基础，然后使用相关分析、回归分析等研究农业总产值与主要农产品的产量、单位面积产量，以及粮食产品的组成部分、水果产品的组成部分、油料作物的组成部分之间的关系；再使用因子分析对主要农产品的产量、单位面积产量等变量提取公因子；最后使用聚类分析依照粮食产品的组成部分、水果产品的组成部分、油料作物的组成部分对各个省市进行聚类，研究各个省市的农产品产出特点。

22.3　数据整理

	下载资源:\video\chap22\…
	下载资源:\sample\chap22\案例22.dta

因为本例采用的是现有数据，所以根据第 1 章介绍的方法直接将所用数据录入 Stata 中即可。我们共设置了 38 个变量，分别是“城市”“农业总产值”“粮食产量”“棉花产量”“油料产量”“麻类产量”“甘蔗产量”“甜菜产量”“烟叶产量”“茶叶产量”“水果产量”“谷物”“稻谷”“小麦”“玉米”“豆类”“薯类”“花生”“油菜籽”“芝麻”“黄红麻”“烤烟”“苹果”“柑桔”“梨”“葡萄”“香蕉”“谷物单位面积产量”“棉花单位面积产量”“花生单位面积产量”“油菜籽单位面积产量”“芝麻单位面积产量”“黄红麻单位面积产量”“甘蔗单位面积产量”“烤烟单位面积产量”“受灾面积（千公顷）”“成灾面积（千公顷）”“甜菜单位面积产量”等。下面把这 38 个变量分别定义为 V1~V38，并分别给这些变量加上标签说明。样本是我国分地区主要农产品产量情况的相关数据。录入完成后数据如图 22.1 所示。

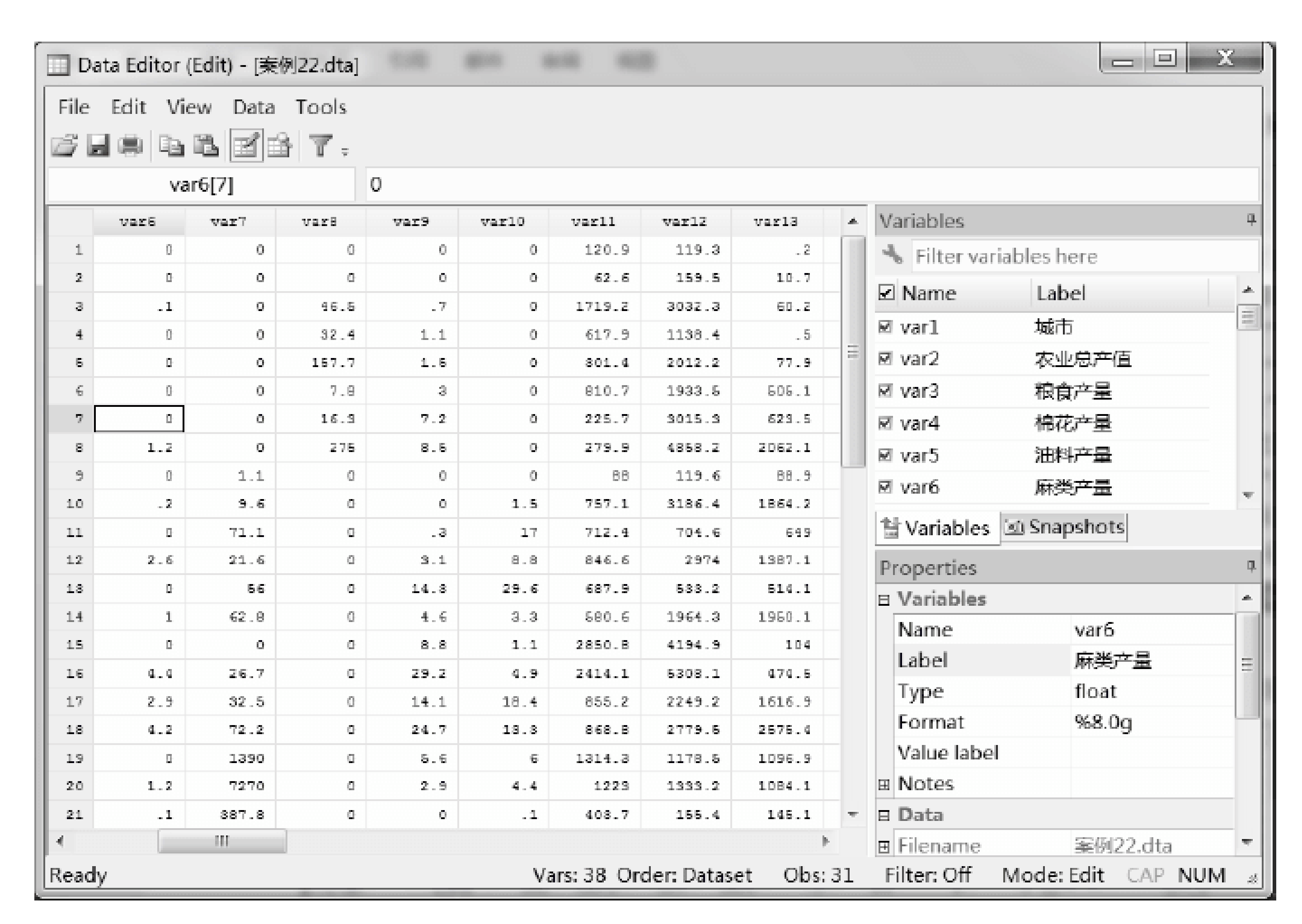


图 22.1　录入数据

先做一下数据保存，然后展开后续分析。

22.4　描述性分析

本案例的数据变量除了城市这一字符串变量外都是定距变量，通过进行定距变量的基本描述性统计，可以得到数据的概要统计指标，包括平均值、最大值、最小值、标准差、百分位数、中位数、偏度系数和峰度系数等。通过获得这些指标，可以从整体上对拟分析的数据进行宏观把握，为后续进行更深入的数据分析做好必要准备。

22.4.1　Stata 分析过程

描述性分析的步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入命令：

```
summarize var2-var38,detail
```

03 设置完毕后，按键盘上的回车键，等待输出结果。

22.4.2 结果分析

在 Stata 14.0 主界面的结果窗口可以看到如图 22.2~图 22.20 所示的分析结果。

```
                          农业总产值
-------------------------------------------------------------
      Percentiles      Smallest
 1%           50             50
 5%          103            103
10%          165            163       Obs                  31
25%          655            165       Sum of Wgt.          31

50%         1136                      Mean           1354.452
                        Largest       Std. Dev.      1010.276
75%         2042           2641
90%         2641           2775       Variance        1020658
95%         3600           3600       Skewness        .754426
99%         3844           3844       Kurtosis       2.932796

                          粮食产量
-------------------------------------------------------------
      Percentiles      Smallest
 1%         93.7           93.7
 5%        103.4          103.4
10%          122          121.8       Obs                  31
25%        672.8            122       Sum of Wgt.          31

50%         1361                      Mean           1842.613
                        Largest       Std. Dev.      1545.331
75%       3135.5         3307.8
90%       3307.8         4426.3       Variance        2388048
95%       5542.5         5542.5       Skewness       .8765954
99%       5570.6         5570.6       Kurtosis        3.05907
```

图 22.2　V2 和 V3 描述性分析结果图

```
                          棉花产量
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0              0
 5%            0              0
10%            0              0       Obs                  31
25%            0              0       Sum of Wgt.          31

50%          1.2                      Mean           21.25161
                        Largest       Std. Dev.      54.03795
75%         22.7           52.6
90%         52.6           65.3       Variance         2920.1
95%         78.5           78.5       Skewness       4.188291
99%        289.8          289.8       Kurtosis       21.08711

                          油料产量
-------------------------------------------------------------
      Percentiles      Smallest
 1%           .7             .7
 5%          1.4            1.4
10%          6.4            1.9       Obs                  31
25%         23.3            6.4       Sum of Wgt.          31

50%         63.5                      Mean           106.6839
                        Largest       Std. Dev.       121.114
75%        141.8          278.4
90%        278.4          304.7       Variance        14668.6
95%          341            341       Skewness       1.832705
99%        532.4          532.4       Kurtosis       6.307122
```

图 22.3　V4 和 V5 描述性分析结果图

```
                          麻类产量
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0              0
 5%            0              0
10%            0              0       Obs                  31
25%            0              0       Sum of Wgt.          31

50%           .1                      Mean           .9516129
                        Largest       Std. Dev.      1.583639
75%          1.2            2.9
90%          2.9            4.2       Variance       2.507914
95%          4.4            4.4       Skewness       1.815864
99%          6.1            6.1       Kurtosis       5.469598

                          甘蔗产量
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0              0
 5%            0              0
10%            0              0       Obs                  31
25%            0              0       Sum of Wgt.          31

50%          1.1                      Mean           369.1452
                        Largest       Std. Dev.      1345.931
75%         62.8          387.8
90%        387.8           1390       Variance        1811531
95%       1898.8         1898.8       Skewness       4.615133
99%         7270           7270       Kurtosis       23.87672
```

图 22.4　V6 和 V7 描述性分析结果图

```
                          甜菜产量
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0              0
 5%            0              0
10%            0              0       Obs                  31
25%            0              0       Sum of Wgt.          31

50%            0                      Mean           34.61613
                        Largest       Std. Dev.      105.9206
75%          7.8           46.5
90%         46.5          157.7       Variance       11219.17
95%          275            275       Skewness       3.661875
99%          519            519       Kurtosis       16.05033

                          烟叶产量
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0              0
 5%            0              0
10%            0              0       Obs                  31
25%           .2              0       Sum of Wgt.          31

50%            3                      Mean           10.10323
                        Largest       Std. Dev.      20.07103
75%          9.4           24.9
90%         24.9           29.2       Variance       402.8463
95%         34.3           34.3       Skewness       3.705151
99%        105.6          105.6       Kurtosis       17.80197
```

图 22.5　V8 和 V9 描述性分析结果图

茶叶产量

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	31
25%	0	0	Sum of Wgt.	31
50%	1.1		Mean	5.235484
		Largest	Std. Dev.	8.090717
75%	6	18.4		
90%	18.4	18.6	Variance	65.4597
95%	23.8	23.8	Skewness	1.62716
99%	29.6	29.6	Kurtosis	4.568942

水果产量

	Percentiles	Smallest		
1%	1.4	1.4		
5%	4.4	4.4		
10%	88	62.6	Obs	31
25%	237.3	88	Sum of Wgt.	31
50%	617.9		Mean	734.4581
		Largest	Std. Dev.	677.8977
75%	868.8	1587.1		
90%	1587.1	1719.2	Variance	459545.3
95%	2414.1	2414.1	Skewness	1.492579
99%	2850.8	2850.8	Kurtosis	5.084141

图 22.6　V10 和 V11 描述性分析结果图

谷物

	Percentiles	Smallest		
1%	59.4	59.4		
5%	91	91		
10%	119.6	119.3	Obs	31
25%	533.2	119.6	Sum of Wgt.	31
50%	1178.5		Mean	1675.468
		Largest	Std. Dev.	1449.734
75%	2779.5	3186.4		
90%	3186.4	4194.9	Variance	2101727
95%	4858.2	4858.2	Skewness	.8734846
99%	5308.1	5308.1	Kurtosis	2.956591

稻谷

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	.5	.2	Obs	31
25%	60.6	.5	Sum of Wgt.	31
50%	474.5		Mean	648.3935
		Largest	Std. Dev.	747.9334
75%	1096.9	1864.2		
90%	1864.2	1950.1	Variance	559404.4
95%	2062.1	2062.1	Skewness	1.056282
99%	2575.4	2575.4	Kurtosis	2.923572

图 22.7　V12 和 V13 描述性分析结果图

小麦

	Percentiles	Smallest		
1%	0	0		
5%	.2	.2		
10%	.8	.3	Obs	31
25%	10.2	.8	Sum of Wgt.	31
50%	54.2		Mean	378.7129
		Largest	Std. Dev.	704.2988
75%	410.9	1215.7		
90%	1215.7	1276.1	Variance	496036.8
95%	2103.9	2103.9	Skewness	2.574964
99%	3123	3123	Kurtosis	9.405153

玉米

	Percentiles	Smallest		
1%	2.8	2.8		
5%	2.8	2.8		
10%	10.5	10.3	Obs	31
25%	78.9	10.5	Sum of Wgt.	31
50%	257		Mean	621.8742
		Largest	Std. Dev.	763.8649
75%	854.6	1696.5		
90%	1696.5	1978.7	Variance	583489.5
95%	2339	2339	Skewness	1.326719
99%	2675.8	2675.8	Kurtosis	3.544054

图 22.8　V14 和 V15 描述性分析结果图

豆类

	Percentiles	Smallest		
1%	1.2	1.2		
5%	1.5	1.5		
10%	2.3	1.7	Obs	31
25%	18.2	2.3	Sum of Wgt.	31
50%	34.8		Mean	61.56452
		Largest	Std. Dev.	104.6299
75%	82.8	115		
90%	115	125.7	Variance	10947.41
95%	171.3	171.3	Skewness	4.069575
99%	577.8	577.8	Kurtosis	20.49131

薯类

	Percentiles	Smallest		
1%	.4	.4		
5%	.6	.6		
10%	1.3	.8	Obs	31
25%	36.9	1.3	Sum of Wgt.	31
50%	67.8		Mean	105.5871
		Largest	Std. Dev.	99.34263
75%	164.2	228.9		
90%	228.9	239.3	Variance	9868.958
95%	284.2	284.2	Skewness	1.467716
99%	441.7	441.7	Kurtosis	5.376823

图 22.9　V16 和 V17 描述性分析结果图

花生

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	.2	0	Obs	31
25%	1.3	.2	Sum of Wgt.	31
50%	9.8		Mean	51.75806
		Largest	Std. Dev.	96.52077
75%	62.7	116.5		
90%	116.5	128.9	Variance	9316.26
95%	338.6	338.6	Skewness	2.882832
99%	429.8	429.8	Kurtosis	10.85243

油菜籽

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	31
25%	.6	0	Sum of Wgt.	31
50%	15.2		Mean	43.30645
		Largest	Std. Dev.	63.45808
75%	66.7	122.8		
90%	122.8	182	Variance	4026.928
95%	214.4	214.4	Skewness	1.708871
99%	220.4	220.4	Kurtosis	4.918908

图 22.10　V18 和 V19 描述性分析结果图

芝麻

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	31
25%	0	0	Sum of Wgt.	31
50%	.2		Mean	1.958065
		Largest	Std. Dev.	4.967077
75%	1.4	3.2		
90%	3.2	6.2	Variance	24.67185
95%	14.6	14.6	Skewness	3.54694
99%	24.1	24.1	Kurtosis	15.0781

黄红麻

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	31
25%	0	0	Sum of Wgt.	31
50%	0		Mean	.2387097
		Largest	Std. Dev.	.8159565
75%	.1	.2		
90%	.2	1.1	Variance	.665785
95%	1.4	1.4	Skewness	4.295561
99%	4.3	4.3	Kurtosis	21.33012

图 22.11　V20 和 V21 描述性分析结果图

烤烟

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	31
25%	0	0	Sum of Wgt.	31
50%	2.6		Mean	9.248387
		Largest	Std. Dev.	19.31317
75%	8.7	23.3		
90%	23.3	29.2	Variance	372.9986
95%	32.5	32.5	Skewness	3.790908
99%	101.8	101.8	Kurtosis	18.3147

苹果

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	31
25%	0	0	Sum of Wgt.	31
50%	10.6		Mean	116.0774
		Largest	Std. Dev.	230.8125
75%	71.5	333.9		
90%	333.9	420.3	Variance	53274.42
95%	837.9	837.9	Skewness	2.431132
99%	902.9	902.9	Kurtosis	8.13174

图 22.12　V22 和 V23 描述性分析结果图

柑桔

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	31
25%	0	0	Sum of Wgt.	31
50%	3.9		Mean	94.96452
		Largest	Std. Dev.	148.433
75%	194.4	355		
90%	355	356.7	Variance	22032.36
95%	378.7	378.7	Skewness	1.149376
99%	420.4	420.4	Kurtosis	2.55378

梨

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	2.9	.1	Obs	31
25%	7.4	2.9	Sum of Wgt.	31
50%	24.2		Mean	50.95161
		Largest	Std. Dev.	77.05201
75%	73	100.5		
90%	100.5	122.7	Variance	5937.012
95%	140.2	140.2	Skewness	3.353506
99%	406.9	406.9	Kurtosis	15.8474

图 22.13　V24 和 V25 描述性分析结果图

葡萄

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	31
25%	6.2	0	Sum of Wgt.	31
50%	14.1		Mean	29.24194
		Largest	Std. Dev.	38.6349
75%	36.4	67.3		
90%	67.3	98.5	Variance	1492.655
95%	112.5	112.5	Skewness	2.291274
99%	175.5	175.5	Kurtosis	8.297456

香蕉

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	31
25%	0	0	Sum of Wgt.	31
50%	0		Mean	33.54516
		Largest	Std. Dev.	87.06618
75%	.2	168.7		
90%	168.7	189.2	Variance	7580.519
95%	205.7	205.7	Skewness	2.767296
99%	384.9	384.9	Kurtosis	10.23758

图 22.14　V26 和 V27 描述性分析结果图

谷物单位面积产量

	Percentiles	Smallest		
1%	3365.7	3365.7		
5%	3735.9	3735.9		
10%	4103.4	3837.1	Obs	31
25%	5001.2	4103.4	Sum of Wgt.	31
50%	5752.3		Mean	5525.152
		Largest	Std. Dev.	997.2727
75%	6199.9	6569.2		
90%	6569.2	6680.1	Variance	994552.9
95%	6821.3	6821.3	Skewness	-.4261357
99%	7581.8	7581.8	Kurtosis	2.653743

棉花单位面积产量

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	31
25%	642.7	0	Sum of Wgt.	31
50%	1076		Mean	1013.826
		Largest	Std. Dev.	610.7311
75%	1489	1769		
90%	1769	1812.3	Variance	372992.5
95%	1821.8	1821.8	Skewness	-.4542562
99%	1940.3	1940.3	Kurtosis	2.212118

图 22.15　V28 和 V29 描述性分析结果图

花生单位面积产量

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	1559.2	1447.8	Obs	31
25%	2426.4	1559.2	Sum of Wgt.	31
50%	2711.9		Mean	2704.184
		Largest	Std. Dev.	1031.113
75%	3558.2	3767.8		
90%	3767.8	4247.7	Variance	1063193
95%	4252.9	4252.9	Skewness	-.8423184
99%	4464.9	4464.9	Kurtosis	4.183686

油菜籽单位面积产量

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	150	0	Obs	31
25%	1181.3	150	Sum of Wgt.	31
50%	1886.8		Mean	1605.952
		Largest	Std. Dev.	760.2736
75%	2142.9	2385		
90%	2385	2525.4	Variance	578015.9
95%	2574.5	2574.5	Skewness	-.8912458
99%	2645.3	2645.3	Kurtosis	2.961897

图 22.16　V30 和 V31 描述性分析结果图

芝麻单位面积产量

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	495.5	0	Obs	31
25%	966.7	495.5	Sum of Wgt.	31
50%	1218.8		Mean	1131.152
		Largest	Std. Dev.	484.9655
75%	1438.5	1665.6		
90%	1665.6	1667.6	Variance	235191.5
95%	1701.6	1701.6	Skewness	-.9919902
99%	1906.7	1906.7	Kurtosis	3.656506

黄红麻单位面积产量

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	31
25%	0	0	Sum of Wgt.	31
50%	666.7		Mean	1800.513
		Largest	Std. Dev.	2157.737
75%	3275.4	4600		
90%	4600	5352.8	Variance	4655828
95%	6700	6700	Skewness	.8812369
99%	6846.7	6846.7	Kurtosis	2.676838

图 22.17　V32 和 V33 描述性分析结果图

甘蔗单位面积产量

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	31
25%	0	0	Sum of Wgt.	31
50%	34902.7		Mean	29647.03
		Largest	Std. Dev.	29417.52
75%	60904.1	64082.4		
90%	64082.4	66599.2	Variance	8.65e+08
95%	67398.2	67398.2	Skewness	.2034883
99%	86735.8	86735.8	Kurtosis	1.493109

烤烟单位面积产量

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	31
25%	714.3	0	Sum of Wgt.	31
50%	2110.9		Mean	1885.842
		Largest	Std. Dev.	1289.332
75%	2622.5	3310		
90%	3310	3428.4	Variance	1662377
95%	4120	4120	Skewness	.0630109
99%	5064.3	5064.3	Kurtosis	2.857195

图 22.18　V34 和 V35 描述性分析结果图

受灾面积(千公顷)

	Percentiles	Smallest		
1%	8.1	8.1		
5%	17.8	17.8		
10%	56.1	24.3	Obs	31
25%	434.7	56.1	Sum of Wgt.	31
50%	1015		Mean	1047.435
		Largest	Std. Dev.	779.2753
75%	1528.2	2117.2		
90%	2117.2	2374.8	Variance	607270
95%	2570.2	2570.2	Skewness	.4518736
99%	2580	2580	Kurtosis	2.173632

成灾面积(千公顷)

	Percentiles	Smallest		
1%	3.2	3.2		
5%	4	4		
10%	18.4	9.2	Obs	31
25%	154	18.4	Sum of Wgt.	31
50%	332.7		Mean	401.3323
		Largest	Std. Dev.	337.7479
75%	682.5	790.3		
90%	790.3	908.7	Variance	114073.6
95%	955	955	Skewness	.863995
99%	1363.7	1363.7	Kurtosis	3.27764

图 22.19　V36 和 V37 描述性分析结果图

甜菜单位面积产量

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	31
25%	0	0	Sum of Wgt.	31
50%	0		Mean	12397.35
		Largest	Std. Dev.	20248.22
75%	32481.8	40221.8		
90%	40221.8	44482	Variance	4.10e+08
95%	57815.9	57815.9	Skewness	1.377929
99%	68799.7	68799.7	Kurtosis	3.615326

图 22.20　V38 描述性分析结果图

从图 22.2~图 22.20 所示的分析结果中可以得到很多信息。此处限于篇幅不再针对各个变量一一展开说明，以变量 V38 为例进行解释。

- 百分位数（Percentiles）：可以看出变量 V38 的第 1 个四分位数（25%）是 0，第 2 个四分位数（50%）是 0。
- 4 个最小值（Smallest）：变量 V38 最小的 4 个数据值分别是 0、0、0、0。
- 4 个最大值（Largest）：变量 V38 最大的 4 个数据值分别是 40221.8、44482、57815.9、68799.7。
- 平均值（Mean）和标准差（Std. Dev）：变量 V38 的平均值为 12397.35，标准差是 20248.22。
- 偏度（Skewness）和峰度（Kurtosis）：变量 V38 的偏度为 1.377929，为正偏度。变量 V38 的峰度为 3.615326，有一个比正态分布略长的尾巴。

从上面的描述性分析结果中可以看出，所有数据中没有极端数据，数据间的量纲差距也在可接受范围之内，可以进入下一步的分析过程。

22.5　相关分析

对于相关分析，准备进行以下几个部分：

- 对“农业总产值”的 9 个来源（“粮食产量”“棉花产量”“油料产量”“麻类产量”“甘蔗产量”“甜菜产量”“烟叶产量”“茶叶产量”“水果产量”）进行简单相关分析。
- 对 9 种农产品的单位面积产量（“谷物单位面积产量”“棉花单位面积产量”“花生单位面积产量”“油菜籽单位面积产量”“芝麻单位面积产量”“黄红麻单位面积产量”“甘蔗单位面积产量”“烤烟单位面积产量”“甜菜单位面积产量”）进行简单相关分析。
- 对“稻谷”“小麦”“玉米”“豆类”“薯类”5 种粮食作物进行简单相关分析。
- 对“花生”“油菜籽”“芝麻”3 种油料作物进行简单相关分析。
- 对“苹果”“柑桔”“梨”“葡萄”“香蕉”5 种水果产品进行简单相关分析。

1. 对“农业总产值”的 10 个来源进行简单相关分析

操作步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入如下命令：

- correlate var3-var11: 本命令旨在使用简单相关分析方法研究 var3~var11 共 9 个变量之间的相关关系。
- pwcorr var3-var11,sidak sig star(0.01): 本命令旨在判断 var3~var11 共 9 个变量之间的相关性在置信水平为 99%时是否显著。

03 设置完毕后，按键盘上的回车键，等待输出结果。

结果分析如图 22.21 和图 22.22 所示。从图 22.21 可以看出，构成“农业总产值”的 10 个来源，大部分变量之间的相关系数不高。

```
. correlate var3-var11
(obs=31)
```

	var3	var4	var5	var6	var7	var8	var9	var10	var11
var3	1.0000								
var4	0.1461	1.0000							
var5	0.7318	0.2104	1.0000						
var6	0.4903	0.2642	0.6630	1.0000					
var7	-0.0713	-0.1077	-0.1072	0.0180	1.0000				
var8	0.1710	0.7631	-0.1086	0.1328	-0.0926	1.0000			
var9	0.1764	-0.0983	0.2113	0.3011	0.1502	-0.1259	1.0000		
var10	0.0420	-0.1056	0.2129	0.3969	0.0996	-0.2182	0.5833	1.0000	
var11	0.5739	0.3562	0.7099	0.2679	0.1373	-0.0030	0.0579	0.0674	1.0000

图 22.21　相关分析结果图 1

```
. pwcorr var3-var11,sidak sig star(0.01)

             |     var3     var4     var5     var6     var7     var8     var9
-------------+---------------------------------------------------------------
        var3 |   1.0000 
             |
             |
        var4 |   0.1461   1.0000 
             |   1.0000
             |
        var5 |   0.7318*  0.2104   1.0000 
             |   0.0001   1.0000
             |
        var6 |   0.4903   0.2642   0.6630*  1.0000 
             |   0.1684   0.9972   0.0017
             |
        var7 |  -0.0713  -0.1077  -0.1072   0.0180   1.0000 
             |   1.0000   1.0000   1.0000   1.0000
             |
        var8 |   0.1710   0.7631* -0.1086   0.1328  -0.0926   1.0000 
             |   1.0000   0.0000   1.0000   1.0000   1.0000
             |
        var9 |   0.1764  -0.0983   0.2113   0.3011   0.1502  -0.1259   1.0000 
             |   1.0000   1.0000   1.0000   0.9772   1.0000   1.0000
             |
       var10 |   0.0420  -0.1056   0.2129   0.3969   0.0996  -0.2182   0.5833 
             |   1.0000   1.0000   1.0000   0.6277   1.0000   0.9999   0.0204
             |
       var11 |   0.5739   0.3562   0.7099*  0.2679   0.1373  -0.0030   0.0579 
             |   0.0262   0.8374   0.0003   0.9965   1.0000   1.0000   1.0000
             |

             |    var10    var11
-------------+------------------
       var10 |   1.0000 
             |
             |
       var11 |   0.0674   1.0000 
             |   1.0000
```

图 22.22 相关分析结果图 2

从图 22.22 中可以看出，“粮食产量”与“油料产量”、“棉花产量”与“甜菜产量”、“油料产量”与“麻类产量”等变量之间的相关性在 1%的显著性水平上显著。

2. 对 9 种农产品的单位面积产量进行简单相关分析

操作步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入命令：

- correlate var28-var35 var38: 本命令旨在使用简单相关分析方法研究 var28~var35、var38 等 9 个变量之间的相关关系。
- pwcorr var28-var35 var38,sidak sig star(0.01)：本命令旨在判断 var28~var35、var38 等 9 个变量之间的相关性在置信水平为 99%时是否显著。

03 设置完毕后，按键盘上的回车键，等待输出结果。

结果分析如图 22.23 和图 22.24 所示。从图 22.23 可以看出，9 种农产品的单位面积产量，大部分变量之间的相关系数不高。

```
. correlate var28-var35 var38
(obs=31)

             |    var28    var29    var30    var31    var32    var33    var34    var35    var38
-------------+---------------------------------------------------------------------------------
       var28 |   1.0000
       var29 |   0.2929   1.0000
       var30 |   0.4496   0.3984   1.0000
       var31 |   0.0903  -0.0914  -0.0129   1.0000
       var32 |   0.5121   0.2738   0.5807  -0.0132   1.0000
       var33 |   0.2113  -0.1396   0.4488   0.0068   0.4005   1.0000
       var34 |   0.0867  -0.0441   0.1137   0.0039   0.2679   0.5318   1.0000
       var35 |  -0.0949   0.0073  -0.1479   0.0782   0.0718  -0.0075  -0.1326   1.0000
       var38 |  -0.0288   0.3098   0.0564   0.0248  -0.0949  -0.4339  -0.5840   0.2104   1.0000
```

图 22.23　相关分析结果图 3

```
. pwcorr var28-var35 var38,sidak sig star(0.01)

             |    var28    var29    var30    var31    var32    var33    var34
-------------+---------------------------------------------------------------
       var28 |   1.0000
             |
             |
       var29 |   0.2929   1.0000
             |   0.9848
             |
       var30 |   0.4496   0.3984   1.0000
             |   0.3326   0.6190
             |
       var31 |   0.0903  -0.0914  -0.0129   1.0000
             |   1.0000   1.0000   1.0000
             |
       var32 |   0.5121   0.2738   0.5807  -0.0132   1.0000
             |   0.1100   0.9948   0.0219   1.0000
             |
       var33 |   0.2113  -0.1396   0.4488   0.0068   0.4005   1.0000
             |   1.0000   1.0000   0.3362   1.0000   0.6064
             |
       var34 |   0.0867  -0.0441   0.1137   0.0039   0.2679   0.5318   1.0000
             |   1.0000   1.0000   1.0000   1.0000   0.9965   0.0721
             |
       var35 |  -0.0949   0.0073  -0.1479   0.0782   0.0718  -0.0075  -0.1326
             |   1.0000   1.0000   1.0000   1.0000   1.0000   1.0000   1.0000
             |
       var38 |  -0.0288   0.3098   0.0564   0.0248  -0.0949  -0.4339  -0.5840
             |   1.0000   0.9663   1.0000   1.0000   1.0000   0.4142   0.0200
             |

             |    var35    var38
-------------+------------------
       var35 |   1.0000
             |
             |
       var38 |   0.2104   1.0000
             |   1.0000
             |
```

图 22.24　相关分析结果图 4

从图 22.24 中可以看出，9 种农产品的单位面积产量等变量之间的相关性都比较差，在 1%的显著性水平上不显著。

3. 对“稻谷”“小麦”“玉米”“豆类”“薯类”5 种粮食作物进行简单相关分析

操作步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入如下命令：

- correlate var13 var14 var15 var16 var17：本命令旨在使用简单相关分析方法研究 var13、var14、var15、var16、var17 共 5 个变量之间的相关关系。
- pwcorr var13 var14 var15 var16 var17,sidak sig star(0.01)：本命令旨在判断 var13、var14、

var15、var16、var17 共 5 个变量之间的相关性在置信水平为 99%时是否显著。

03 设置完毕后，按键盘上的回车键，等待输出结果。

结果分析如图 22.25 和图 22.26 所示。从图 22.25 可以看出，“稻谷”“小麦”“玉米”“豆类”“薯类”共 5 种粮食作物之间的相关系数不大。

从图 22.26 中可以看出，仅有“玉米”与“豆类”之间的相关性在 1%的显著性水平上显著。

```
. correlate  var13 var14 var15 var16 var17
(obs=31)

             |   var13    var14    var15    var16    var17
-------------+---------------------------------------------
       var13 |  1.0000
       var14 | -0.0354   1.0000
       var15 |  0.0241   0.4232   1.0000
       var16 |  0.4209   0.0669   0.6491   1.0000
       var17 |  0.1877   0.1221   0.2192   0.2191   1.0000
```

图 22.25 相关分析结果图 5

```
. pwcorr var13 var14 var15 var16 var17,sidak sig star(0.01)

             |   var13    var14    var15    var16    var17
-------------+---------------------------------------------
       var13 |  1.0000
             |
       var14 | -0.0354   1.0000
             |  1.0000
       var15 |  0.0241   0.4232   1.0000
             |  1.0000   0.1634
       var16 |  0.4209   0.0669   0.6491*  1.0000
             |  0.1693   1.0000   0.0008
       var17 |  0.1877   0.1221   0.2192   0.2191   1.0000
             |  0.9762   0.9992   0.9323   0.9326
```

图 22.26 相关分析结果图 6

4. 对“花生”“油菜籽”“芝麻”3 种油料作物进行简单相关分析

操作步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入如下命令：

- correlate var18 var19 var20：本命令旨在使用简单相关分析方法研究 var18、var19、var20 共 3 个变量之间的相关关系。
- pwcorr var18 var19 var20,sidak sig star(0.01)：本命令旨在判断 var18、var19、var20 共 3 个变量之间的相关性在置信水平为 99%时是否显著。

03 设置完毕后，按键盘上的回车键，等待输出结果。

结果分析如图 22.27 和图 22.28 所示。从图 22.27 可以看出，“花生”“油菜籽”“芝麻”共 3 种油料作物之间的相关系数不大。

从图 22.28 中可以看出，仅有“花生”与“芝麻”之间的相关性在 1%的显著性水平上显著。

```
. correlate  var18 var19 var20
(obs=31)

             |   var18    var19    var20
-------------+---------------------------
       var18 |  1.0000
       var19 |  0.1003   1.0000
       var20 |  0.6508   0.4375   1.0000
```

图 22.27 相关分析结果图 7

```
. pwcorr  var18 var19 var20,sidak sig star(0.01)

             |   var18    var19    var20
-------------+---------------------------
       var18 |  1.0000
             |
       var19 |  0.1003   1.0000
             |  0.9317
       var20 |  0.6508*  0.4375   1.0000
             |  0.0002   0.0410
```

图 22.28 相关分析结果图 8

5. 对“苹果”“柑桔”“梨”“葡萄”“香蕉”5 种水果产品进行简单相关分析

操作步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入如下命令：

- correlate var23 var24 var25 var26 var27: 本命令旨在使用简单相关分析方法研究 var23、var24、var25、var26、var27 共 5 个变量之间的相关关系。
- pwcorr var23 var24 var25 var26 var27,sidak sig star(0.01): 本命令旨在判断 var23、var24、var25、var26、var27 共 5 个变量之间的相关性在置信水平为 99%时是否显著。

03 设置完毕后，按键盘上的回车键，等待输出结果。

结果分析如图 22.29 和图 22.30 所示。从图 22.29 可以看出，“苹果”“柑桔”“梨”“葡萄”“香蕉”5 种水果产品之间的相关系数不大。

从图 22.30 中可以看出，仅有“梨”与“葡萄”变量之间的相关性在 1%的显著性水平上显著。

```
. correlate var23 var24 var25 var26 var27
(obs=31)

             |    var23    var24    var25    var26    var27
-------------+---------------------------------------------
       var23 |   1.0000
       var24 |  -0.2845   1.0000
       var25 |   0.4499  -0.1701   1.0000
       var26 |   0.4145  -0.2288   0.6220   1.0000
       var27 |  -0.1929   0.4019  -0.1836  -0.1758   1.0000
```

图 22.29 相关分析结果图 9

```
. pwcorr   var23 var24 var25 var26 var27,sidak sig star(0.01)

             |    var23    var24    var25    var26    var27
-------------+---------------------------------------------
       var23 |   1.0000
             |
             |
       var24 |  -0.2845   1.0000
             |   0.7243
             |
       var25 |   0.4499  -0.1701   1.0000
             |   0.1056   0.9885
             |
       var26 |   0.4145  -0.2288   0.6220*  1.0000
             |   0.1864   0.9120   0.0019
             |
       var27 |  -0.1929   0.4019  -0.1836  -0.1758   1.0000
             |   0.9712   0.2239   0.9797   0.9853
```

图 22.30 相关分析结果图 10

22.6 回归分析

对于回归分析，准备进行以下几个部分：

- 以“农业总产值”为因变量，以农业为自变量，进行最小二乘线性回归。
- 以“农业总产值”为因变量，以“谷物单位面积产量”“棉花单位面积产量”“花生单位面积产量”“油菜籽单位面积产量”“芝麻单位面积产量”“黄红麻单位面积产量”“甘蔗单位面积产量”“烤烟单位面积产量”“甜菜单位面积产量”“受灾面积（千公顷）”“成灾面积（千公顷）”为自变量，进行最小二乘线性回归。

1. 以“农业总产值”为因变量，以“粮食产量”“棉花产量”“油料产量”“麻类产量”“甘蔗产量”“甜菜产量”“烟叶产量”“茶叶产量”“水果产量”“受灾面积（千公顷）”“成灾面积（千公顷）”为自变量，进行最小二乘回归

建立线性模型：

Var2=a*Var3+b*Var4+c*Var5+d*Var6+e*Var7+f*Var8+g*Var9+h*Var10+i*Var11 + u

普通最小二乘回归分析步骤及结果如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入如下命令：

- sw regress var2 var3-var11,pr(0.1): 本命令的含义是使用逐步回归分析方法，以“农业总产值”为因变量，以“粮食产量”“棉花产量”“油料产量”“麻类产量”“甘蔗产量”“甜菜产量”“烟叶产量”“茶叶产量”“水果产量”为自变量，进行最小二乘回归分析。
- predict yhat: 本命令旨在获得因变量的拟合值。
- predict e,resid: 本命令旨在获得回归模型的估计残差。
- rvfplot: 本命令旨在绘制残差与回归得到的拟合值的散点图，探索数据是否存在异方差。
- estat imtest,white: 本命令为怀特检验，旨在检验数据是否存在异方差。
- estat hettest,iid: 本命令为 BP 检验，旨在使用得到的拟合值来检验数据是否存在异方差。
- estat hettest,rhs iid: 本命令为 BP 检验，旨在使用方程右边的解释数据来检验变量是否存在异方差。

03 设置完毕后，按键盘上的回车键进行确认。

在 Stata 14.0 主界面的结果窗口可以看到如图 22.31~图 22.37 所示的分析结果。

图 22.31 是使用逐步回归分析方法，以“农业总产值”为因变量，以“粮食产量”“棉花产量”“油料产量”“麻类产量”“甘蔗产量”“甜菜产量”“烟叶产量”“茶叶产量”“水果产量”为自变量，进行最小二乘回归分析的结果。

```
. sw regress  var2   var3-var11,pr(0.1)
                      begin with full model
p = 0.9613 >= 0.1000  removing var6
p = 0.6571 >= 0.1000  removing var5
p = 0.4521 >= 0.1000  removing var7
p = 0.2370 >= 0.1000  removing var9

      Source |       SS           df       MS      Number of obs   =        31
-------------+----------------------------------   F(  5,      25) =     73.94
       Model |  28680238.6         5  5736047.72   Prob > F        =    0.0000
    Residual |  1939493.06        25  77579.7224   R-squared       =    0.9367
-------------+----------------------------------   Adj R-squared   =    0.9240
       Total |  30619731.7        30  1020657.72   Root MSE        =    278.53

------------------------------------------------------------------------------
        var2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        var3 |    .376556   .0460278     8.18   0.000     .2817599    .4713521
        var4 |   4.158882   1.961544     2.12   0.044      .119006    8.198758
       var10 |   18.14509   6.500415     2.79   0.010     4.757231    31.53294
       var11 |    .644364   .1238177     5.20   0.000     .3893567    .8993714
        var8 |  -1.792251   .9692901    -1.85   0.076    -3.788541    .2040392
       _cons |   66.00589   89.92312     0.73   0.470    -119.1942     251.206
------------------------------------------------------------------------------
```

图 22.31　回归分析结果图 1

从上述分析结果中可以看出共有 31 个样本参与了分析，模型的 F 值(5, 25) =73.94，P 值（Prob > F）= 0.0000，说明模型整体上是非常显著的。模型的可决系数（R-squared）为 0.9367，模型修正的可决系数（Adj R-squared）为 0.9240，说明模型的解释能力是非常优秀且接近完美的。

模型经过 4 次剔除变量后得到最终结果。第 1 个模型是包含全部自变量的全模型，该模

型中 var6 变量的系数显著性 P 值高达 0.9613，被剔除掉；第 2 个模型是剔除掉自变量 var6 以后的模型，该模型中 var5 变量的系数显著性 P 值高达 0.6571，被剔除掉；第 3 个模型是剔除掉自变量 var6、var5 以后的模型，该模型中 var7 变量的系数显著性 P 值高达 0.4521，被剔除掉；第 4 个模型是剔除掉自变量 var6、var5、var7 以后的模型，该模型中 var9 变量的系数显著性 P 值高达 0.2370，被剔除掉。剔除自变量 var6、var5、var7、var9 以后得到最终回归模型。

在最终回归模型中，变量 var3 的系数标准误是 0.0460278，t 值为 8.18，P 值为 0.000，系数是非常显著的，95%的置信区间为[0.2817599, 0.4713521]。变量 var4 的系数标准误是 1.961544，t 值为 2.12，P 值为 0.044，系数是非常显著的，95%的置信区间为[0.119006, 8.198758]。变量 var10 的系数标准误是 6.500415，t 值为 2.79，P 值为 0.010，系数是非常显著的，95%的置信区间为[4.757231, 31.53294]。变量 var11 的系数标准误是 0.1238177，t 值为 5.20，P 值为 0.000，系数是非常显著的，95%的置信区间为[0.3893567, 0.8993714]。变量 var8 的系数标准误是 0.9692901，t 值为-1.85，P 值为 0.076，系数是比较显著的，95%的置信区间为[-3.788541, 0.2040392]。常数项的系数标准误是 89.92312，t 值为 0.73，P 值为 0.470，系数是非常不显著的，95%的置信区间为[-119.1942, 251.206]。

最终最小二乘回归模型的方程是：

$$var2=0.376556*var3+4.158882*var4-1.792251*var8+18.14509*var10+0.644364*var11+66.00589$$

图 22.32 是对因变量的拟合值的预测。

	var31	var32	var33	var34	var35	var36	var37	var38	yhat
1	150	966.7	0	0	0	56.1	18.4	0	190.1899
2	0	1438.5	0	0	0	8.1	3.2	0	197.2138
3	1436.2	1402	2219.4	0	1816.2	1383.3	528.1	37242.9	2556.693
4	991.2	1018.3	0	0	3428.4	1015	544.8	57815.9	881.5217
5	1098.7	495.5	0	0	4120	2036.6	908.7	40221.8	877.4384
6	2071	1906.7	0	0	2733.3	450.2	173.1	44482	1341.308
7	0	1398.1	0	0	2745.1	616.4	221.7	32481.8	1381.275
8	2574.5	1385.1	0	0	2440.4	1536.8	682.5	33526.3	1851.137
9	2150.3	928.6	0	64082.4	0	24.3	9.2	0	170.7292
10	2385	1667.6	4600	58940.5	1700	1032.1	332.7	0	1929.368
11	1958	1665.6	3596.4	62764.4	0	431.1	159.3	0	1141.142
12	1917.2	1298.9	2997.9	39924.7	2701.8	1317.2	198.5	0	2109.098
13	1369	1252.3	3275.4	60904.1	2110.9	133.1	48.3	0	1299.705
14	1228.4	997.3	4490.9	44891.1	2294.8	1075.3	426.7	0	1332.469
15	2525.4	1701.6	6700	0	2622.5	2117.2	415.6	0	3916.14
16	2016.4	1360	5352.8	67398.2	2345	1477.6	380.2	0	3956.407
17	1930.9	1636.2	3430	41651.7	2024.3	2580	790.3	0	2069.097
18	1558.9	1451.5	3096.6	49813.7	2353.3	2374.8	955	0	2068.414
19	1181.3	1189.3	2365.6	86735.8	2301.9	501.7	118.3	0	1534.257
20	1037.8	1218.8	2570.9	66599.2	1734.2	1437.9	638.1	0	1473.171
21	0	988.1	6846.7	64072.7	714.3	516.6	198	0	398.7427
22	1791	995.9	1390.4	34902.7	1951.3	816	280.5	0	709.3965
23	2223.2	1301.1	2216.2	46945	2037.4	1528.2	720.7	17020.4	2149.09
24	1468.5	1041.9	666.7	36409.9	1625.1	2570.2	1363.7	4365.9	584.3438

图 22.32　回归分析结果图 2

因变量预测拟合值是根据自变量的值和得到的回归方程计算出来的，主要用于预测未来。在图 22.32 中可以看到 yhat 的值与 var2 的值是比较相近的，所以拟合的回归模型还是不错的。

图 22.33 是回归分析得到的残差序列。

	var32	var33	var34	var35	var36	var37	var38	yhat	e
1	966.7	0	0	0	56.1	18.4	0	190.1899	-27.18991
2	1438.5	0	0	0	8.1	3.2	0	197.2138	-17.21379
3	1402	2219.4	0	1816.2	1383.3	528.1	37242.9	2556.693	218.3066
4	1018.3	0	0	3428.4	1015	544.8	57815.9	881.5217	-114.5217
5	495.5	0	0	4120	2036.6	908.7	40221.8	877.4384	180.5616
6	1906.7	0	0	2733.3	450.2	173.1	44482	1341.308	-34.30784
7	1398.1	0	0	2745.1	616.4	221.7	32481.8	1381.275	-361.2748
8	1385.1	0	0	2440.4	1536.8	682.5	33526.3	1851.137	-49.13711
9	928.6	0	64082.4	0	24.3	9.2	0	170.7292	-5.729195
10	1667.6	4600	58940.5	1700	1032.1	332.7	0	1929.368	711.6322
11	1665.6	3596.4	62764.4	0	431.1	159.3	0	1141.142	10.85814
12	1298.9	2997.9	39924.7	2701.8	1317.2	198.5	0	2109.098	-394.0982
13	1252.3	3275.4	60904.1	2110.9	133.1	48.3	0	1299.705	-163.7053
14	997.3	4490.9	44891.1	2294.8	1075.3	426.7	0	1332.469	-414.4686
15	1701.6	6700	0	2622.5	2117.2	415.6	0	3916.14	-72.14037
16	1360	5352.8	67398.2	2345	1477.6	380.2	0	3956.407	-356.4069
17	1636.2	3430	41651.7	2024.3	2580	790.3	0	2069.097	229.9033
18	1451.5	3096.6	49813.7	2353.3	2374.8	955	0	2068.414	323.5858
19	1189.3	2365.6	86735.8	2301.9	501.7	118.3	0	1534.257	507.7433
20	1218.8	2570.9	66599.2	1734.2	1437.9	638.1	0	1473.171	128.8293
21	988.1	6846.7	64072.7	714.3	516.6	198	0	398.7427	2.257314
22	995.9	1390.4	34902.7	1951.3	816	280.5	0	709.3965	41.60347
23	1301.1	2216.2	46945	2037.4	1528.2	720.7	17020.4	2149.09	304.9101
24	1041.9	666.7	36409.9	1625.1	2570.2	1363.7	4365.9	584.3438	70.65617

图 22.33　回归分析结果图 3

图 22.34 是我们上面几步得到的残差与得到的拟合值的散点图。

从图 22.34 中可以看出，残差并没有随着拟合值的大小的不同而不同，而是围绕 0 值上下随机波动的，所以数据很可能是不存在异方差的。

图 22.35 是怀特检验的检验结果。

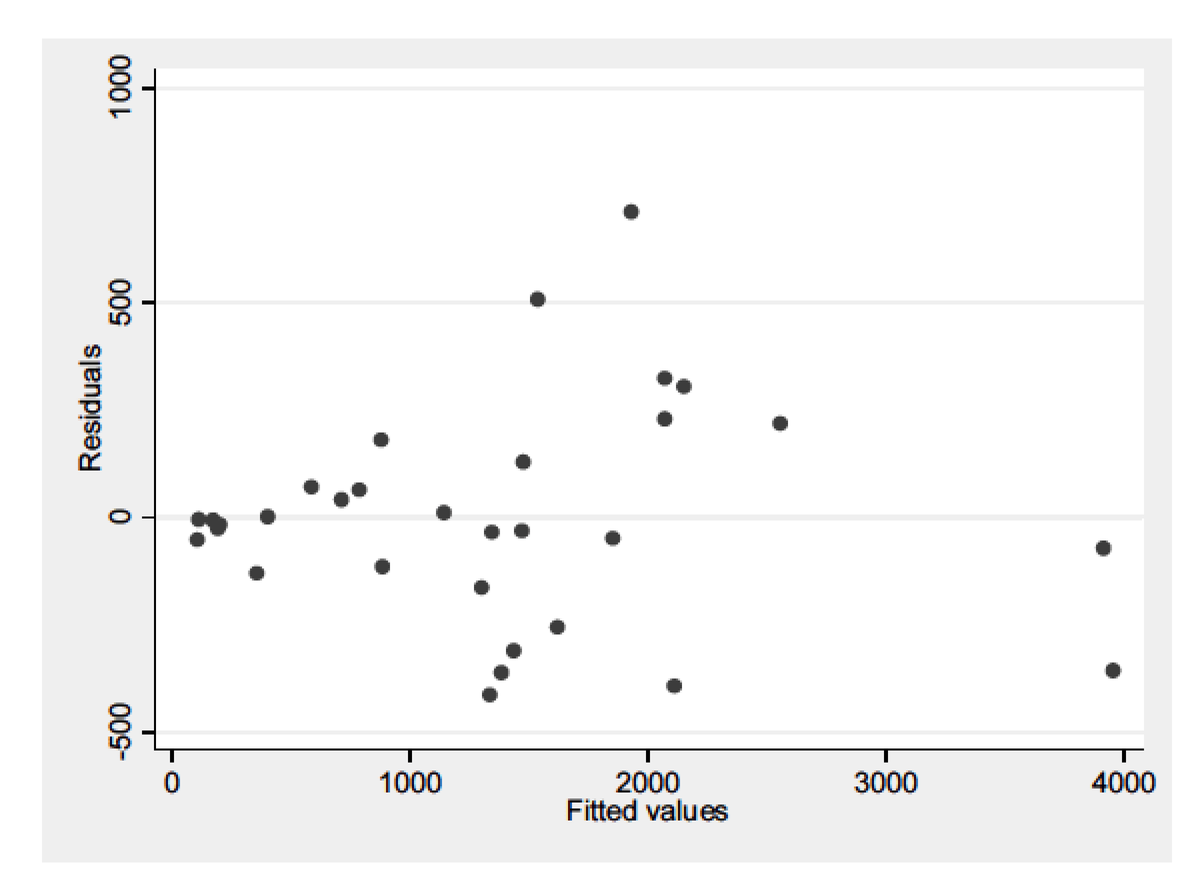
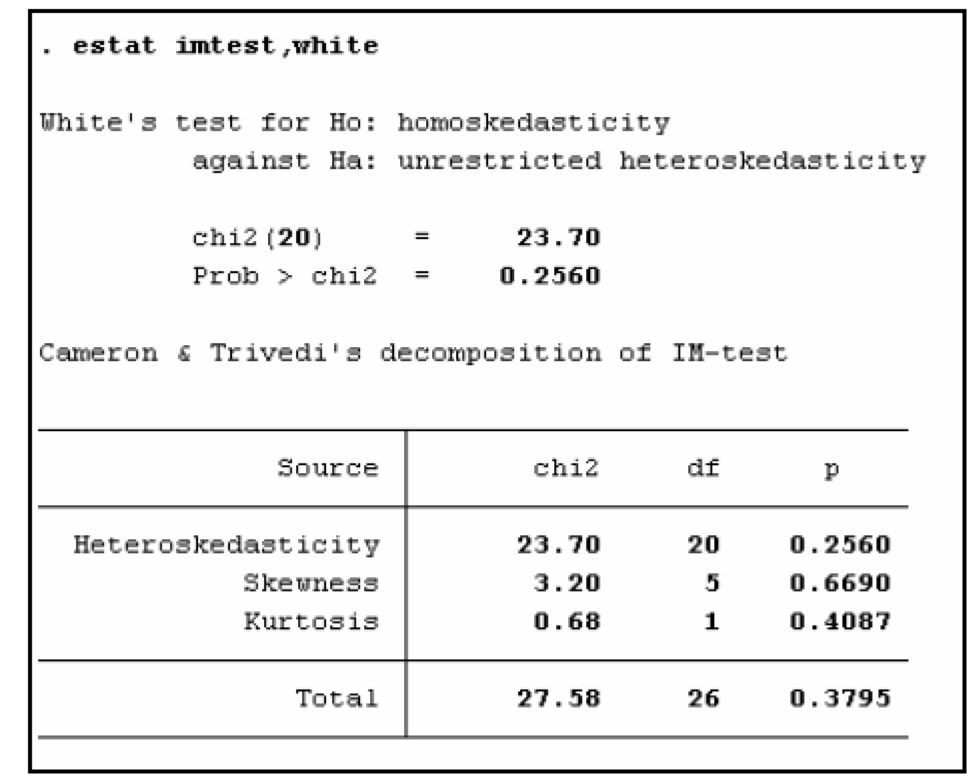


图 22.34　回归分析结果图 4



```
. estat imtest,white

White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(20)      =     23.70
         Prob > chi2   =    0.2560

Cameron & Trivedi's decomposition of IM-test

              Source |       chi2     df      p
---------------------+-----------------------------
  Heteroskedasticity |      23.70     20    0.2560
            Skewness |       3.20      5    0.6690
            Kurtosis |       0.68      1    0.4087
---------------------+-----------------------------
               Total |      27.58     26    0.3795
```

图 22.35　回归分析结果图 5

怀特检验的原假设是数据为同方差。从图 22.35 中可以看出，P 值为 0.2560，非常显著地接受了同方差的原假设，认为不存在异方差。

图 22.36~图 22.37 是 BP 检验的检验结果。其中图 22.36 是使用得到的拟合值对数据进行异方差检验的结果，图 22.37 是使用方程右边的解释变量对数据进行异方差检验的结果。

```
. estat hettest,iid

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of var2

         chi2(1)      =     3.40
         Prob > chi2  =   0.0651
```

图 22.36　回归分析结果图 6

```
. estat hettest,rhs iid

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: var3 var4 var10 var11 var8

         chi2(5)      =     7.44
         Prob > chi2  =   0.1902
```

图 22.37　回归分析结果图 7

BP 检验的原假设是数据为同方差。从图 22.36 和图 22.37 中可以看出，P 值均大于 0.05，非常显著地接受了同方差的原假设，认为不存在异方差，所以没有必要使用稳健的标准差进行回归。

经过以上最小二乘回归分析可以发现我国农业总产值水平与“粮食产量”“棉花产量”“甜菜产量”“茶叶产量”以及“水果产量”都有一定的显著关系。具体而言，“粮食产量”“棉花产量”“茶叶产量”以及“水果产量”有拉动效应，尤其是茶叶产量，每增加一个单位会带来对应农业总产值的 18 倍多的增加；甜菜产量对农业总产值水平有拖后效应，在一定程度上说明种植这种作物是不经济的。

2. 以“农业总产值”为因变量，以“谷物单位面积产量”“棉花单位面积产量”“花生单位面积产量”“油菜籽单位面积产量”“芝麻单位面积产量”“黄红麻单位面积产量”“甘蔗单位面积产量”“烤烟单位面积产量”“甜菜单位面积产量”“受灾面积（千公顷）”“成灾面积（千公顷）”为自变量，进行最小二乘线性回归

建立线性模型：

var2=a*var3+b*var4+c*var5+d*var6+e*var7+f*var8+g*var9+h*var10+i*var11 + u

普通最小二乘回归分析的步骤及结果如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中输入如下命令。

- sw regress var2 var28 var29 var30 var31 var32 var33 var34 var35 var36 var37 var38,pr(0.1)：本命令的含义是使用逐步回归分析方法，以“谷物单位面积产量”“棉花单位面积产量”“花生单位面积产量”“油菜籽单位面积产量”“芝麻单位面积产量”“黄红麻单位面积产量”“甘蔗单位面积产量”“烤烟单位面积产量”“甜菜单位面积产量”“受灾面积（千公顷）”“成灾面积（千公顷）”为自变量，进行最小二乘回归分析。
- predict yhat：本命令旨在获得因变量的拟合值。
- predict e,resid：本命令旨在获得回归模型的估计残差。
- rvfplot：本命令旨在绘制残差与回归得到的拟合值的散点图，探索数据是否存在异方差。
- estat imtest,white：本命令为怀特检验，旨在检验数据是否存在异方差。
- estat hettest,iid：本命令为 BP 检验，旨在使用得到的拟合值来检验数据是否存在异方差。
- estat hettest,rhs iid：本命令为 BP 检验，旨在使用方程右边的解释数据来检验变量是否存在异方差。

03 设置完毕后，按键盘上的回车键进行确认。

在 Stata 14.0 主界面的结果窗口可以看到如图 22.38~图 22.44 所示的分析结果。

图 22.38 是使用逐步回归分析方法，以“谷物单位面积产量”“棉花单位面积产量”“花生单位面积产量”“油菜籽单位面积产量”“芝麻单位面积产量”“黄红麻单位面积产量”“甘蔗单位面积产量”“烤烟单位面积产量”“甜菜单位面积产量”“受灾面积（千公顷）”“成灾面积（千公顷）”为自变量，进行最小二乘回归分析的结果。

```
. sw regress  var2    var28 var29 var30 var31 var32 var33 var34 var35 var36 var37 var38,pr(0.1)
                      begin with full model
p = 0.8016 >= 0.1000  removing var28
p = 0.3663 >= 0.1000  removing var35
p = 0.3307 >= 0.1000  removing var34
p = 0.2565 >= 0.1000  removing var29
p = 0.1851 >= 0.1000  removing var37
p = 0.2168 >= 0.1000  removing var38

      Source |       SS       df       MS              Number of obs =      31
-------------+------------------------------           F(  5,    25) =   17.55
       Model |  23831225.6     5  4766245.13           Prob > F      =  0.0000
    Residual |  6788506.03    25  271540.241           R-squared     =  0.7783
-------------+------------------------------           Adj R-squared =  0.7340
       Total |  30619731.7    30  1020657.72           Root MSE      =   521.1

------------------------------------------------------------------------------
        var2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       var33 |   .1263766   .0514549     2.46   0.021     .0204033      .23235
       var36 |    .472352   .1306429     3.62   0.001     .2032879    .7414161
       var30 |   .2965623   .1187119     2.50   0.019     .0520707     .541054
       var31 |   .3739334   .1282699     2.92   0.007     .1097566    .6381101
       var32 |   .4717766   .2478286     1.90   0.069     -.038636    .9821892
       _cons |  -1303.978    358.614    -3.64   0.001    -2042.558    -565.399
------------------------------------------------------------------------------
```

图 22.38　回归分析结果图 8

从上述分析结果中可以看出共有 31 个样本参与了分析，模型的 F 值(5, 25) =17.55，P 值（Prob > F） = 0.0000，说明模型整体上是非常显著的。模型的可决系数（R-squared）为 0.7783，模型修正的可决系数（Adj R-squared）为 0.7340，说明模型的解释能力是比较不错的。

模型经过 6 次剔除变量后得到最终结果。第 1 个模型是包含全部自变量的全模型，该模型中 var28 变量的系数显著性 P 值高达 0.8016，被剔除掉；第 2 个模型是剔除掉自变量 var28 以后的模型，该模型中 var35 变量的系数显著性 P 值高达 0.3663，被剔除掉；第 3 个模型是剔除掉自变量 var28、var35 以后的模型，该模型中 var34 变量的系数显著性 P 值高达 0.3307，被剔除掉；第 4 个模型是剔除掉自变量 var28、var35、var34 以后的模型，该模型中 var29 变量的系数显著性 P 值高达 0.2565，被剔除掉；第 5 个模型是剔除掉自变量 var28、var35、var34、var29 以后的模型，该模型中 var37 变量的系数显著性 P 值高达 0.1851，被剔除掉；第 6 个模型是剔除掉自变量 var28、var35、var34、var29、var37 以后的模型，该模型中 var38 变量的系数显著性 P 值高达 0.2168，被剔除掉。剔除掉自变量 var28、var35、var34、var29、var37、var38 以后，我们得到最终回归模型。

在最终回归模型中，变量 var33 的系数标准误是 0.0514549，t 值为 2.46，P 值为 0.021，系数是非常显著的，95%的置信区间为[0.0204033, 0.23235]。变量 var36 的系数标准误是 0.1306429，t 值为 3.62，P 值为 0.001，系数是非常显著的，95%的置信区间为[0.2032879, 0.7414161]。变量 var30 的系数标准误是 0.1187119，t 值为 2.50，P 值为 0.019，系数是非常显著的，95%的置信区间为[0.0520707, 0.541054]。变量 var31 的系数标准误是 0.1282699，t 值为 2.92，P 值为 0.007，系数是非常显著的，95%的置信区间为[0.1097566, 0.6381101]。变量 var32

的系数标准误是 0.2478286，t 值为 1.90，P 值为 0.069，系数是比较显著的，95%的置信区间为[-0.038636, 0.9821892]。常数项的系数标准误是 358.614，t 值为-3.64，P 值为 0.001，系数是非常显著的，95%的置信区间为[-2042.558, -565.399]。

最终最小二乘回归模型的方程是：

var2=0.1263766*var33+0.472352 *var36+0.2965623*var30 +0.3739334 *var31
+0.4717766* var32 -1303.978

图 22.39 是对因变量拟合值的预测。

	var31	var32	var33	var34	var35	var36	var37	var38	yhat
1	150	966.7	0	0	0	56.1	18.4	0	120.2123
2	0	1438.5	0	0	0	8.1	3.2	0	433.7265
3	1436.2	1402	2219.4	0	1816.2	1383.3	528.1	37242.9	1889.747
4	991.2	1018.3	0	0	3428.4	1015	544.8	57815.9	751.2806
5	1098.7	495.5	0	0	4120	2036.6	908.7	40221.8	818.312
6	2071	1906.7	0	0	2733.3	450.2	173.1	44482	1499.153
7	0	1398.1	0	0	2745.1	616.4	221.7	32481.8	548.4979
8	2574.5	1385.1	0	0	2440.4	1536.8	682.5	33526.3	1791.557
9	2150.3	928.6	0	64082.4	0	24.3	9.2	0	739.9992
10	2385	1667.6	4600	58940.5	1700	1032.1	332.7	0	2538.343
11	1958	1665.6	3596.4	62764.4	0	431.1	159.3	0	1708.353
12	1917.2	1298.9	2997.9	39924.7	2701.8	1317.2	198.5	0	2350.885
13	1369	1252.3	3275.4	60904.1	2110.9	133.1	48.3	0	1041.033
14	1228.4	997.3	4490.9	44891.1	2294.8	1075.3	426.7	0	1523.193
15	2525.4	1701.6	6700	0	2622.5	2117.2	415.6	0	3549.623
16	2016.4	1360	5352.8	67398.2	2345	1477.6	380.2	0	2727.303
17	1930.9	1636.2	3430	41651.7	2024.3	2580	790.3	0	2902.973
18	1558.9	1451.5	3096.6	49813.7	2353.3	2374.8	955	0	2274.177
19	1181.3	1189.3	2365.6	86735.8	2301.9	501.7	118.3	0	1040.351
20	1037.8	1218.8	2570.9	66599.2	1734.2	1437.9	638.1	0	1447.388
21	0	988.1	6846.7	64072.7	714.3	516.6	198	0	1020.373
22	1791	995.9	1390.4	34902.7	1951.3	816	280.5	0	991.9919
23	2223.2	1301.1	2216.2	46945	2037.4	1528.2	720.7	17020.4	1862.682
24	1468.5	1041.9	666.7	36409.9	1625.1	2570.2	1363.7	4365.9	1497.381

图 22.39　回归分析结果图 9

因变量预测拟合值是根据自变量的值和得到的回归方程计算出来的，主要用于预测未来。在图 22.39 中可以看到 yhat 的值与 var2 的值是比较相近的，所以拟合的回归模型还是不错的。

图 22.40 是回归分析得到的残差序列。

	var32	var33	var34	var35	var36	var37	var38	yhat	e
1	966.7	0	0	0	56.1	18.4	0	120.2123	42.78773
2	1438.5	0	0	0	8.1	3.2	0	433.7265	-253.7265
3	1402	2219.4	0	1816.2	1383.3	528.1	37242.9	1889.747	885.2526
4	1018.3	0	0	3428.4	1015	544.8	57815.9	751.2806	15.71944
5	495.5	0	0	4120	2036.6	908.7	40221.8	818.312	239.688
6	1906.7	0	0	2733.3	450.2	173.1	44482	1499.153	-192.1529
7	1398.1	0	0	2745.1	616.4	221.7	32481.8	548.4979	471.5021
8	1385.1	0	0	2440.4	1536.8	682.5	33526.3	1791.557	10.44254
9	928.6	0	64082.4	0	24.3	9.2	0	739.9992	-574.9992
10	1667.6	4600	58940.5	1700	1032.1	332.7	0	2538.343	102.6573
11	1665.6	3596.4	62764.4	0	431.1	159.3	0	1708.353	-556.3528
12	1298.9	2997.9	39924.7	2701.8	1317.2	198.5	0	2350.885	-635.8851
13	1252.3	3275.4	60904.1	2110.9	133.1	48.3	0	1041.033	94.96683
14	997.3	4490.9	44891.1	2294.8	1075.3	426.7	0	1523.193	-605.1925
15	1701.6	6700	0	2622.5	2117.2	415.6	0	3549.623	294.3768
16	1360	5352.8	67398.2	2345	1477.6	380.2	0	2727.303	872.6967
17	1636.2	3430	41651.7	2024.3	2580	790.3	0	2902.973	-603.9734
18	1451.5	3096.6	49813.7	2353.3	2374.8	955	0	2274.177	117.8231
19	1189.3	2365.6	86735.8	2301.9	501.7	118.3	0	1040.351	1001.649
20	1218.8	2570.9	66599.2	1734.2	1437.9	638.1	0	1447.388	154.6124
21	988.1	6846.7	64072.7	714.3	516.6	198	0	1020.373	-619.3731
22	995.9	1390.4	34902.7	1951.3	816	280.5	0	991.9919	-240.992
23	1301.1	2216.2	46945	2037.4	1528.2	720.7	17020.4	1862.682	591.3181
24	1041.9	666.7	36409.9	1625.1	2570.2	1363.7	4365.9	1497.381	-842.3813

图 22.40　回归分析结果图 10

图 22.41 是上面几步得到的残差与得到的拟合值的散点图。

从图 22.41 中可以看出，残差并没有随着拟合值的大小的不同而不同，而是围绕 0 值上下随机波动的，所以，数据很可能是不存在异方差的。

图 22.42 是怀特检验的检验结果。

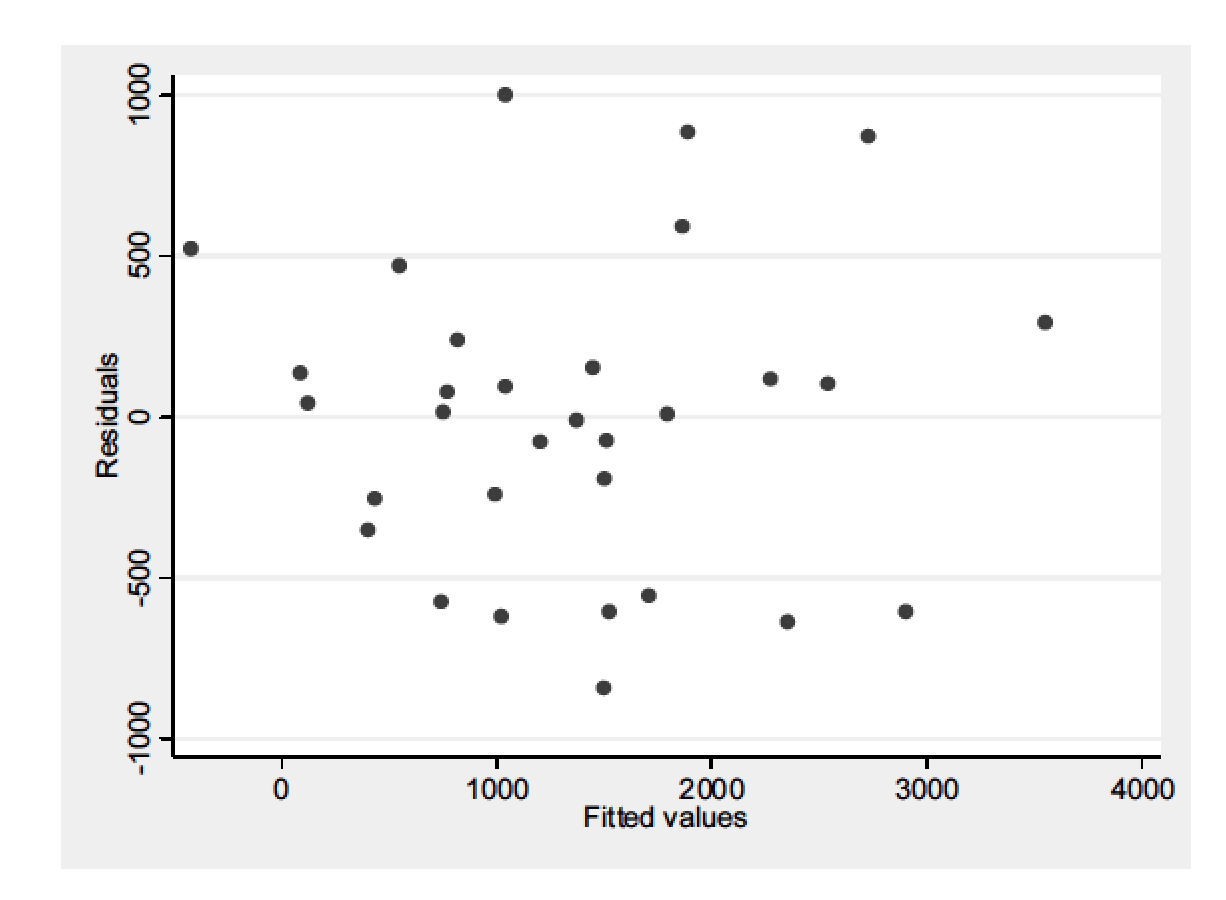
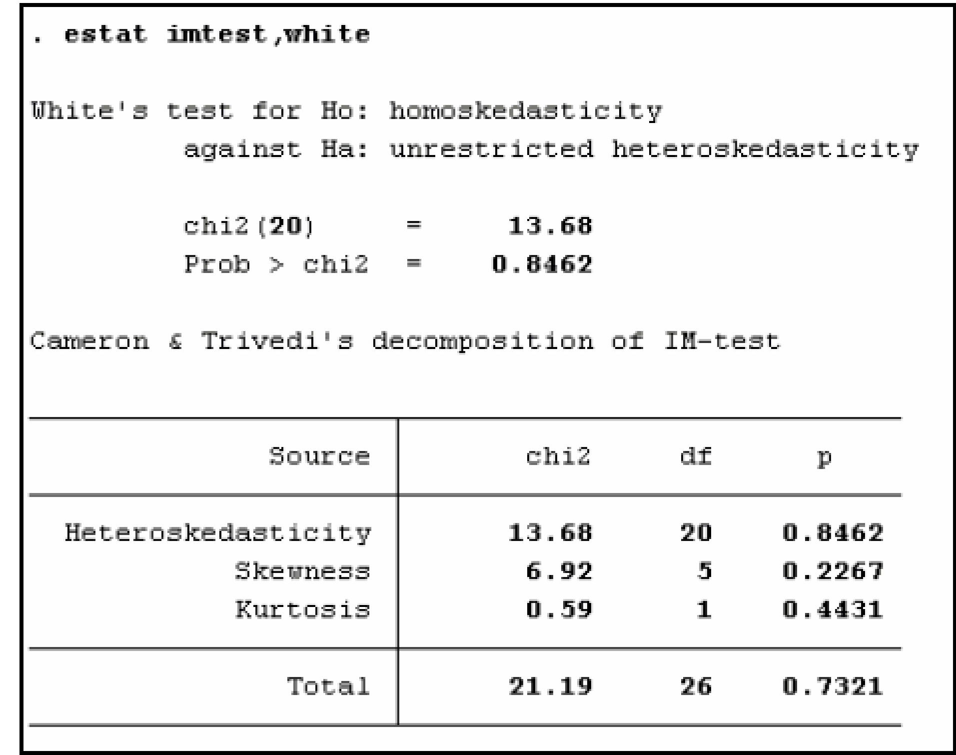


图 22.41　回归分析结果图 11



```
. estat imtest,white

White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(20)      =     13.68
         Prob > chi2   =    0.8462

Cameron & Trivedi's decomposition of IM-test
```

Source	chi2	df	p
Heteroskedasticity	13.68	20	0.8462
Skewness	6.92	5	0.2267
Kurtosis	0.59	1	0.4431
Total	21.19	26	0.7321

图 22.42　回归分析结果图 12

怀特检验的原假设是数据为同方差。从图 22.42 中可以看出，P 值为 0.8462，非常显著地接受了同方差的原假设，认为不存在异方差。

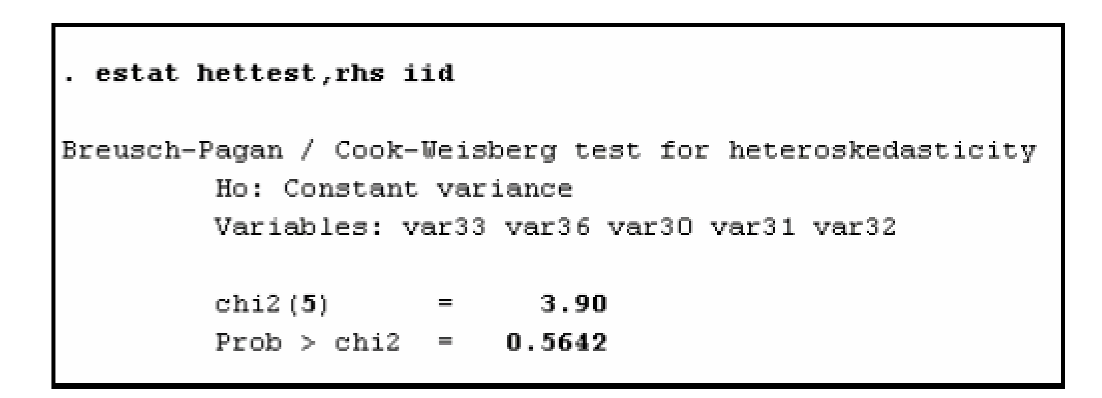
图 22.43 和图 22.44 是 BP 检验的检验结果。其中，图 22.43 是使用得到的拟合值对数据进行异方差检验的结果，图 22.44 是使用方程右边的解释变量对数据进行异方差检验的结果。

```
. estat hettest,iid

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of var2

         chi2(1)      =     1.33
         Prob > chi2  =   0.2486
```

图 22.43　回归分析结果图 13



```
. estat hettest,rhs iid

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: var33 var36 var30 var31 var32

         chi2(5)      =     3.90
         Prob > chi2  =   0.5642
```

图 22.44　回归分析结果图 14

BP 检验的原假设是数据为同方差。从图 22.43 和图 22.44 中可以看出，P 值均大于 0.05，非常显著地接受了同方差的原假设，认为不存在异方差，所以没有必要使用稳健的标准差进行回归。

经过以上最小二乘回归分析可以发现我国农业总产值水平与“花生单位面积产量”“油菜籽单位面积产量”“芝麻单位面积产量”“黄红麻单位面积产量”以及“受灾面积（千公顷）”都有一定的显著关系。具体而言，这些变量都对我国的农业总产值有显著拉动效应。“花生单位面积产量”“油菜籽单位面积产量”“芝麻单位面积产量”“黄红麻单位面积产量”对我国的农业总产值有显著拉动效应，说明这些作物都是经济的，也就是量的提高能够带来价值的提高，“受灾面积（千公顷）”对我国的农业总产值有显著拉动效应，说明“谷贱伤农”的道理在我国是存在的，受灾面积的扩大会带来产量的降低，但这却能带来价格的提高，而且价格提高的幅度要更大，造成总价值也会提高。

22.7 因子分析

对于因子分析，准备从以下两部分进行：

- 对“粮食产量”“棉花产量”“油料产量”“麻类产量”“甘蔗产量”“甜菜产量”“烟叶产量”“茶叶产量”“水果产量”9种农产品产量变量提取公因子。
- 对“谷物单位面积产量”“棉花单位面积产量”“花生单位面积产量”“油菜籽单位面积产量”“芝麻单位面积产量”“黄红麻单位面积产量”“甘蔗单位面积产量”“烤烟单位面积产量”“甜菜单位面积产量”9种作物单位面积产量提取公因子。

1. 对“粮食产量”“棉花产量”“油料产量”“麻类产量”“甘蔗产量”“甜菜产量”“烟叶产量”“茶叶产量”“水果产量”9种农产品产量变量提取公因子

操作步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中分别输入如下命令并按键盘上的回车键进行确认。

- factor var3-var11,pcf：本命令的含义是采用主成分因子法对变量 V3~V11 进行因子分析。
- rotate：本命令的含义是采用最大方差正交旋转法对因子结构进行旋转。
- loadingplot,factors(2) yline(0) xline(0)：本命令的含义是绘制因子旋转后的因子载荷图。
- predict f1 f2 f3 f4：本命令的含义是展示因子分析后各个样本的因子得分情况。
- correlate f1 f2 f3 f4：本命令的含义是展示系统提取的4个主因子的相关系数矩阵。
- scoreplot,mlabel(var1) yline(0) xline(0)：本命令的含义是展示每个样本的因子得分示意图。
- estat kmo：本命令的含义是展示本例因子分析的 KMO 检验结果。
- screeplot：本命令的含义是展示本例因子分析所提取的各个因子的特征值碎石图。

03 设置完毕后，等待输出结果。

在 Stata 14.0 主界面的结果窗口可以看到如图 22.45~图 22.53 所示的分析结果。
图 22.45 展示的是因子分析的基本情况。

```
. factor var3-var11,pcf
(obs=31)

Factor analysis/correlation                        Number of obs    =       31
    Method: principal-component factors            Retained factors =        4
    Rotation: (unrotated)                          Number of params =       30

    --------------------------------------------------------------------------
         Factor  |   Eigenvalue   Difference        Proportion   Cumulative
    -------------+------------------------------------------------------------
        Factor1  |      3.03466      0.99734            0.3372       0.3372
        Factor2  |      2.03732      0.76071            0.2264       0.5636
        Factor3  |      1.27661      0.24494            0.1418       0.7054
        Factor4  |      1.03167      0.43893            0.1146       0.8200
        Factor5  |      0.59274      0.03324            0.0659       0.8859
        Factor6  |      0.55950      0.24044            0.0622       0.9481
        Factor7  |      0.31906      0.22779            0.0355       0.9835
        Factor8  |      0.09127      0.03411            0.0101       0.9936
        Factor9  |      0.05716            .            0.0064       1.0000
    --------------------------------------------------------------------------
    LR test: independent vs. saturated:  chi2(36) =  146.83 Prob>chi2 = 0.0000

Factor loadings (pattern matrix) and unique variances

    -------------------------------------------------------------------
        Variable |  Factor1   Factor2   Factor3   Factor4 |   Uniqueness
    -------------+----------------------------------------+------------
            var3 |   0.7991    0.0942   -0.2816   -0.0513 |      0.2706
            var4 |   0.4026    0.7413    0.3957    0.1003 |      0.1216
            var5 |   0.9031   -0.0716   -0.3141   -0.1011 |      0.0703
            var6 |   0.7698   -0.1014    0.2438   -0.1509 |      0.3149
            var7 |  -0.0032   -0.2788    0.1363    0.9212 |      0.0551
            var8 |   0.1585    0.7858    0.5177    0.0419 |      0.0876
            var9 |   0.3697   -0.5872    0.4814   -0.0268 |      0.2861
           var10 |   0.3534   -0.6398    0.4809   -0.1173 |      0.2207
           var11 |   0.7341    0.1180   -0.3648    0.3482 |      0.1928
    -------------------------------------------------------------------
```

图 22.45　因子分析结果图 1

图 22.45 的上半部分说明是因子分析模型的一般情况，从图中可以看出共有 31 个样本（Number of obs= 31）参与了分析，提取保留的因子共有 4 个（Retained factors = 4），模型 LR 检验的卡方值（LR test: independent vs. saturated: chi2(36)）为 146.83，P 值（Prob>chi2）为 0.0000，模型非常显著。图 22.45 的上半部分最左列（Factor）说明的是因子名称，可以看出模型共提取了 9 个因子。Eigenvalue 列表示的是提取因子的特征值情况，只有前 4 个因子的特征值是大于 1 的，其中第 1 个因子的特征值是 3.03466，第 2 个因子的特征值是 2.03732。Proportion 列表示的是提取因子的方差贡献率，其中第 1 个因子的方差贡献率为 33.72%，第 2 个因子的方差贡献率为 22.64%。Cumulative 列表示的是提取因子的累计方差贡献率，其中前两个因子的累计方差贡献率为 56.36%。

图 22.45 的下半部分说明的是模型的因子载荷矩阵以及变量的未被解释部分。其中 Variable 列表示的是变量名称，Factor1、Factor2、Factor3、Factor4 这 4 列分别说明的是提取的前 4 个主因子（特征值大于 1 的）对各个变量的解释程度，本例中，Factor1 主要解释的是 V3、V5、V6、V11 这 4 个变量的信息，Factor2 主要解释的是 V4、V8 变量的信息，Factor3 主要解释的是 V9、V10 这 2 个变量的信息，Factor4 主要解释的是 V7 变量的信息。Uniqueness 列表示变量未被提取的前 4 个主因子解释的部分，可以发现在舍弃其他主因子的情况下，信息的损失量是比较小的。

图 22.46 展示的是对因子结构进行旋转的结果。经研究表明，旋转操作有助于进一步简化因子结构。Stata 14.0 支持的旋转方式有两种：一种是最大方差正交旋转，一般适用于相互独立的因子或者成分，也是系统默认的情况；另一种是 Promax 斜交旋转，它允许因子或者成分之

间存在相关关系。此处我们选择系统默认方式，当然后面的操作也证明了这种方式的恰当性。

```
. rotate

Factor analysis/correlation                        Number of obs    =        31
    Method: principal-component factors           Retained factors =         4
    Rotation: orthogonal varimax (Kaiser off)      Number of params =        30

    --------------------------------------------------------------------------
         Factor  |     Variance   Difference        Proportion   Cumulative
    -------------+------------------------------------------------------------
        Factor1  |      2.66253      0.82391            0.2958       0.2958
        Factor2  |      1.83862      0.03007            0.2043       0.5001
        Factor3  |      1.80854      0.73797            0.2009       0.7011
        Factor4  |      1.07058            .            0.1190       0.8200
    --------------------------------------------------------------------------
    LR test: independent vs. saturated:  chi2(36)  =  146.83 Prob>chi2 = 0.0000

Rotated factor loadings (pattern matrix) and unique variances

    ---------------------------------------------------------------------
        Variable |  Factor1   Factor2   Factor3   Factor4 |   Uniqueness
    -------------+----------------------------------------+--------------
            var3 |   0.8384    0.1076    0.0705   -0.0998 |      0.2706
            var4 |   0.2145    0.9110   -0.0486   -0.0062 |      0.1216
            var5 |   0.9356   -0.0172    0.1986   -0.1206 |      0.0703
            var6 |   0.5594    0.2438    0.5487   -0.1079 |      0.3149
            var7 |  -0.0224   -0.0510    0.0899    0.9663 |      0.0551
            var8 |  -0.0592    0.9485   -0.0769   -0.0585 |      0.0876
            var9 |   0.0831   -0.0829    0.8264    0.1313 |      0.2861
           var10 |   0.0613   -0.1362    0.8684    0.0529 |      0.2207
           var11 |   0.8430    0.0986   -0.0982    0.2779 |      0.1928
    ---------------------------------------------------------------------

Factor rotation matrix

    --------------------------------------------------------
                 |  Factor1   Factor2   Factor3   Factor4
    -------------+------------------------------------------
         Factor1 |   0.8877    0.2534    0.3843   -0.0051
         Factor2 |   0.0420    0.7702   -0.6073   -0.1904
         Factor3 |  -0.4547    0.5779    0.6705    0.0986
         Factor4 |   0.0587    0.0931   -0.1841    0.9767
    --------------------------------------------------------
```

图 22.46　因子分析结果图 2

图 22.46 包括 3 部分内容，第 1 部分说明的是因子旋转模型的一般情况，从图中可以看出共有 31 个样本（Number of obs = 31）参与了分析，提取保留的因子共有 4 个（Retained factors = 4），模型 LR 检验的卡方值（LR test: independent vs. saturated: chi2(36)）为 146.83，P 值（Prob>chi2）为 0.0000，模型非常显著。最左列（Factor）说明的是因子名称，可以看出模型旋转后共提取了 4 个因子。Proportion 列表示的是提取因子的方差贡献率，其中第 1 个因子的方差贡献率为 29.58%，第 2 个因子的方差贡献率为 20.43%。Cumulative 列表示的是提取因子的累计方差贡献率，其中前两个因子的累计方差贡献率为 50.01%。

图 22.46 的第 2 部分说明的是模型的因子载荷矩阵以及变量的未被解释部分。其中 Variable 列表示的是变量名称，Factor1、Factor2 两列分别说明的是旋转提取的两个主因子对各个变量的解释程度，本例中，Factor1 主要解释的是 V3、V5、V6、V11 这 4 个变量的信息，Factor2 主要解释的是 V4、V8 变量的信息，Factor3 主要解释的是 V6、V9、V10 这 3 个变量的信息，Factor4 主要解释的是 V7 这个变量的信息。Uniqueness 列表示变量未被提取的前 4 个主因子解释的部分，可以发现在舍弃其他主因子的情况下，信息的损失量是很小的。

图 22.46 的第 3 部分展示的是因子旋转矩阵的一般情况，提取的 4 个因子相关关系不明显。

图 22.47 展示的是因子旋转后的因子载荷图。因子载荷图可以使用户更加直观地看出各个变量被前两个因子的解释情况。

与前面的分析相同， Factor1 主要解释的是 V3、V5、V6、V11 这 4 个变量的信息，Factor2

主要解释的是 V4、V8 变量的信息。

图 22.48 展示的是因子分析后各个样本的因子得分情况。因子得分的概念是通过将每个变量标准化为平均数等于 0 和方差等于 1，然后以因子分析系数进行加权合计为每个因子构成的线性情况。以因子的方差贡献率为权数对因子进行加权求和，即可得到每个样本的因子综合得分。

根据图 22.48 展示的因子得分系数矩阵，可以写出各公因子的表达式。值得一提的是，在表达式中各个变量已经不是原始变量而是标准化变量。

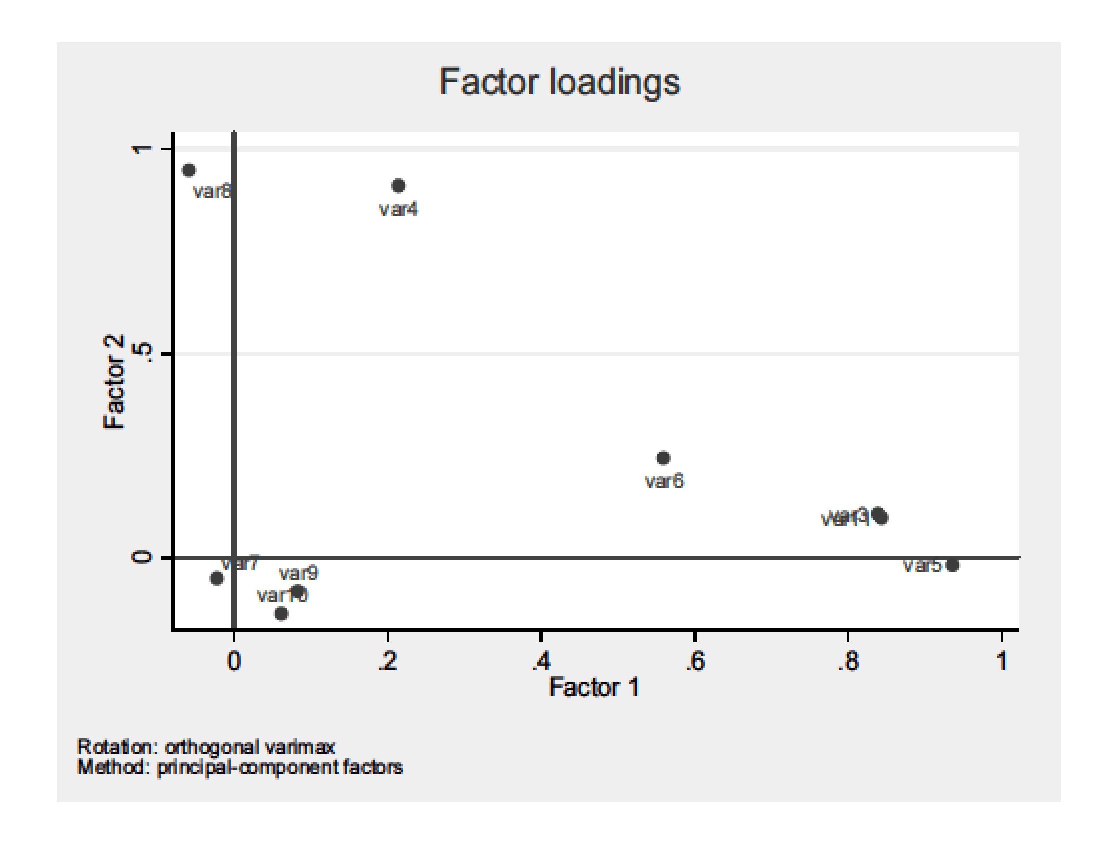


图 22.47 因子分析结果图 3

```
. predict f1 f2 f3 f4
(regression scoring assumed)

Scoring coefficients (method = regression; based on varimax rotated factors)

    Variable    Factor1   Factor2   Factor3   Factor4

        var3    0.33310  -0.02975  -0.06566  -0.08043
        var4   -0.00218   0.50206   0.01996   0.05557
        var5    0.36885  -0.10295  -0.01123  -0.11476
        var6    0.12768   0.12268   0.28270  -0.11582
        var7   -0.00278   0.03914  -0.01013   0.90870
        var8   -0.11944   0.54843   0.05025   0.00596
        var9   -0.07692   0.02436   0.47946   0.06613
       var10   -0.08776  -0.00527   0.50899  -0.01468
       var11    0.36692  -0.02777  -0.19594   0.28918
```

图 22.48 因子分析结果图 4

表达式如下：

F1= 0.33310*粮食产量 -0.00218*棉花产量+ 0.36885*油料产量+ 0.12768*麻类产量 -0.00278*甘蔗产量-0.11944*甜菜产量-0.07692*烟叶产量-0.08776*茶叶产量+ 0.36692*水果产量

F2= -0.02975*粮食产量+ 0.50206*棉花产量-0.10295*油料产量+ 0.12268*麻类产量+ 0.03914*甘蔗产量+ 0.54843*甜菜产量+ 0.02436*烟叶产量-0.00527*茶叶产量-0.02777 *水果产量

F3= -0.06566*粮食产量+ 0.01996 *棉花产量-0.01123*油料产量+ 0.28270*麻类产量 -0.01013 *甘蔗产量+ 0.05025*甜菜产量+ 0.47946*烟叶产量+ 0.50899*茶叶产量 -0.19594*水果产量

F4= -0.08043*粮食产量+ 0.05557*棉花产量-0.11476*油料产量-0.11582*麻类产量+ 0.90870*甘蔗产量+ 0.00596*甜菜产量+ 0.06613*烟叶产量-0.01468*茶叶产量+ 0.28918*水果产量

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 22.49 所示的因子得分数据。

	var36	var37	var38	yhat	e	f1	f2	f3	f4
1	56.1	18.4	0	120.2123	42.78773	-.9642293	-.3213035	-.5018237	-.2995308
2	8.1	3.2	0	433.7265	-253.7265	-.9895813	-.2531257	-.4839852	-.3185186
3	1383.3	528.1	37242.9	1889.747	885.2526	.9363884	.2902701	-1.025731	.1551169
4	1015	544.8	57815.9	751.2806	15.71944	-.4526328	-.1502929	-.6486573	-.1478412
5	2036.6	908.7	40221.8	818.312	239.688	-.1582025	.3343371	-.5518637	-.4520874
6	450.2	173.1	44482	1499.153	-192.1529	.1619328	-.443021	-.7181299	-.2067387
7	616.4	221.7	32481.8	548.4979	471.5021	-.088558	-.3389218	-.4878626	-.4523824
8	1536.8	682.5	33526.3	1791.557	10.44254	.1171138	1.0749	-.2336315	-.5804495
9	24.3	9.2	0	739.9992	-574.9992	-.9804896	-.3166365	-.4922297	-.3128957
10	1032.1	332.7	0	2538.343	102.6573	.5003065	-.2866548	-.6952333	-.314745
11	431.1	159.3	0	1708.353	-556.3528	-.5704197	-.3707918	.3728636	-.0966848
12	1317.2	198.5	0	2350.885	-635.8851	.825759	-.0392532	.2462311	-.49062
13	133.1	48.3	0	1041.033	94.96683	-.835061	-.3785313	1.511759	-.0799373
14	1075.3	426.7	0	1523.193	-605.1925	.0690226	-.2580256	-.2263728	-.3172107
15	2117.2	415.6	0	3549.623	294.3768	2.426652	-.0663071	-1.196892	.4268113
16	1477.6	380.2	0	2727.303	872.6967	3.250701	-.2430997	.3609014	-.2838193
17	2580	790.3	0	2902.973	-603.9734	.823513	.0656744	1.192705	-.5147316
18	2374.8	955	0	2274.177	117.8231	.797987	-.0292438	1.326751	-.5077395
19	501.7	118.3	0	1040.351	1001.649	.1350352	-.4285859	-.4070666	1.005225
20	1437.9	638.1	0	1447.388	154.6124	.08545	-.1270487	-.3748358	4.87859
21	516.6	198	0	1020.373	-619.3731	-.7648228	-.323356	-.5659779	.0638317
22	816	280.5	0	991.9919	-240.992	-.4878839	-.2672773	.0612537	-.4034041
23	1528.2	720.7	17020.4	1862.682	591.3181	1.111581	-.1368615	2.002157	-.7844226
24	2570.2	1363.7	4365.9	1497.381	-842.3813	-.747983	-.3551414	.6586224	-.2846191

图 22.49　因子分析结果图 5

图 22.50 展示的是系统提取的 4 个主因子的相关系数矩阵。

```
. correlate f1 f2 f3 f4
(obs=31)

             |       f1       f2       f3       f4
-------------+------------------------------------
          f1 |   1.0000
          f2 |   0.0000   1.0000
          f3 |  -0.0000  -0.0000   1.0000
          f4 |  -0.0000   0.0000   0.0000   1.0000
```

图 22.50　因子分析结果图 6

从图 22.50 中可以看出，提取的 4 个主因子之间几乎没有什么相关关系，这也说明了在前面对因子进行旋转的操作环节中采用最大方差正交旋转方式是明智的。值得说明的是，图中有的相关系数是-0.0000 并非是不正确的，这是因为 Stata 14.0 只保留了 4 位小数所导致，例如真实的数据有可能是-0.00001，那么结果显示的就是-0.0000。

图 22.51 展示的是每个样本在前两个主因子维度上的因子得分示意图。

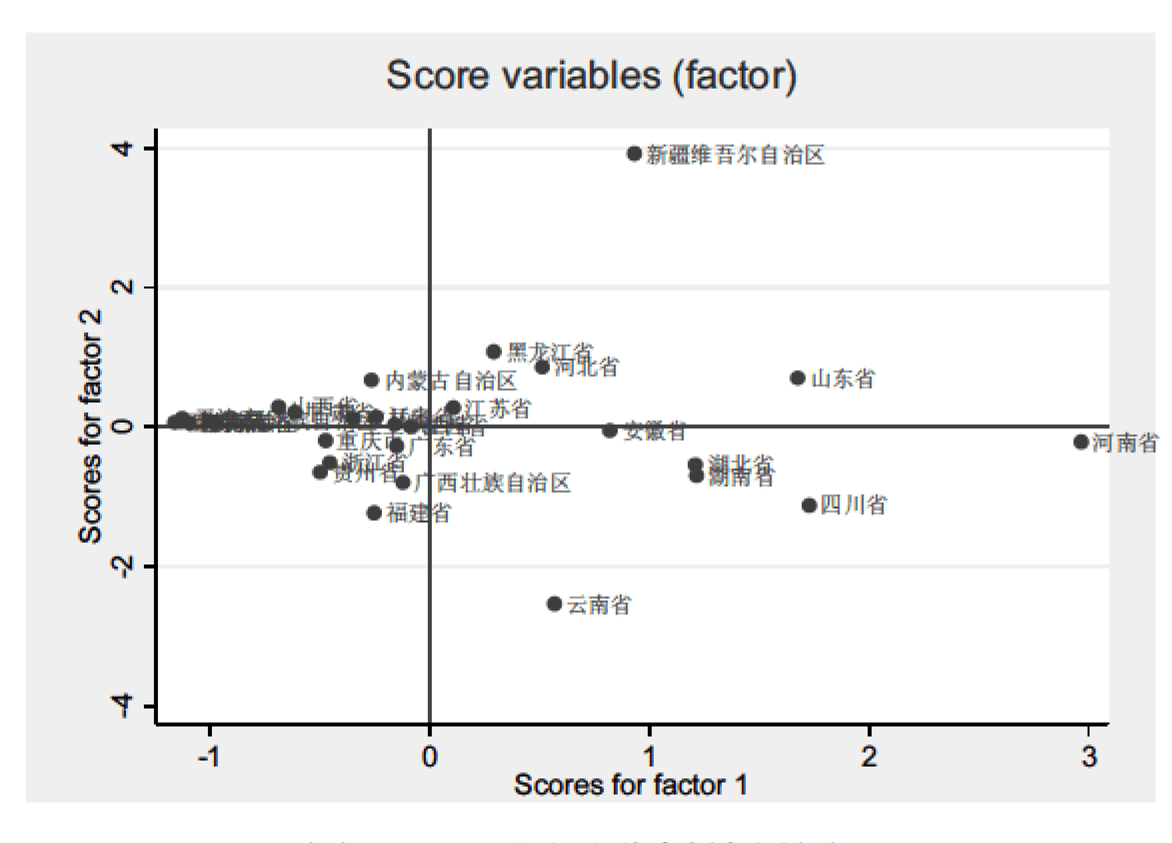


图 22.51　因子分析结果图 7

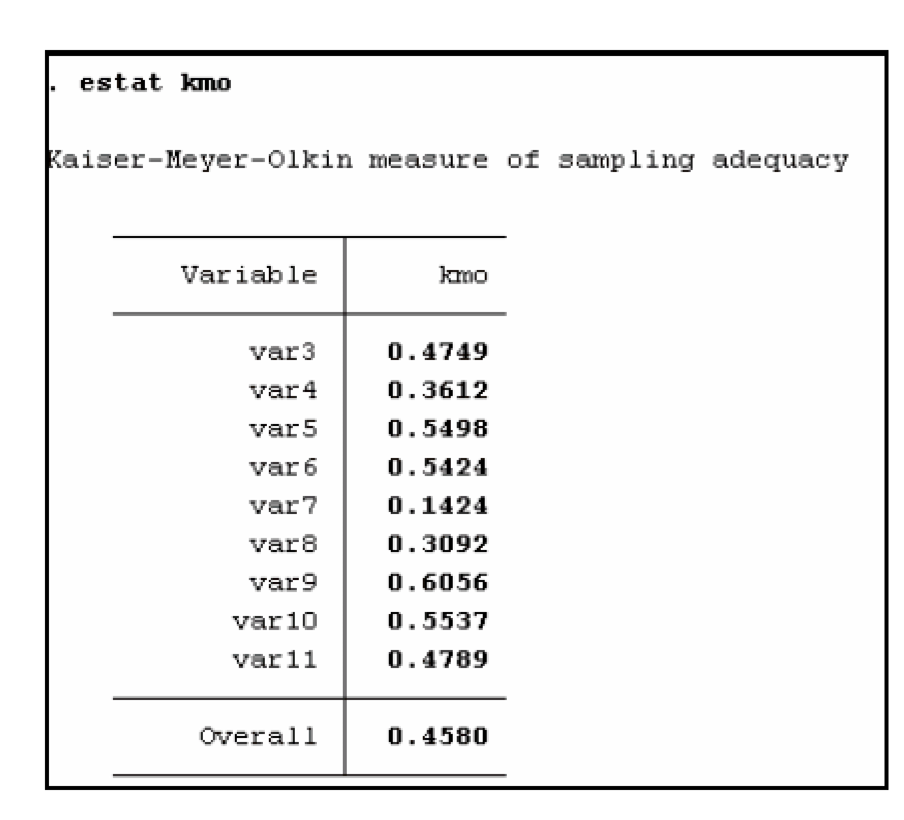
从图 22.51 中可以看出，所有的样本被分到 4 个象限，可以比较直观地看出各个样本的因

子得分分布情况。

图 22.52 展示的是本例因子分析的 KMO 检验结果。

KMO 检验是为了看数据是否适合进行因子分析，其取值范围是 0~1。其中，0.9~1 表示极好、0.8~0.9 表示可奖励的、0.7~0.8 表示还好、0.6~0.7 表示中等。本例中总体（Overall）KMO 的取值为 0.4580，表明因子分析的效果是差强人意的。

图 22.53 展示的是本例因子分析所提取的各个因子的特征值碎石图。



```
. estat kmo

Kaiser-Meyer-Olkin measure of sampling adequacy
```

Variable	kmo
var3	0.4749
var4	0.3612
var5	0.5498
var6	0.5424
var7	0.1424
var8	0.3092
var9	0.6056
var10	0.5537
var11	0.4789
Overall	0.4580


图 22.52　因子分析结果图 8

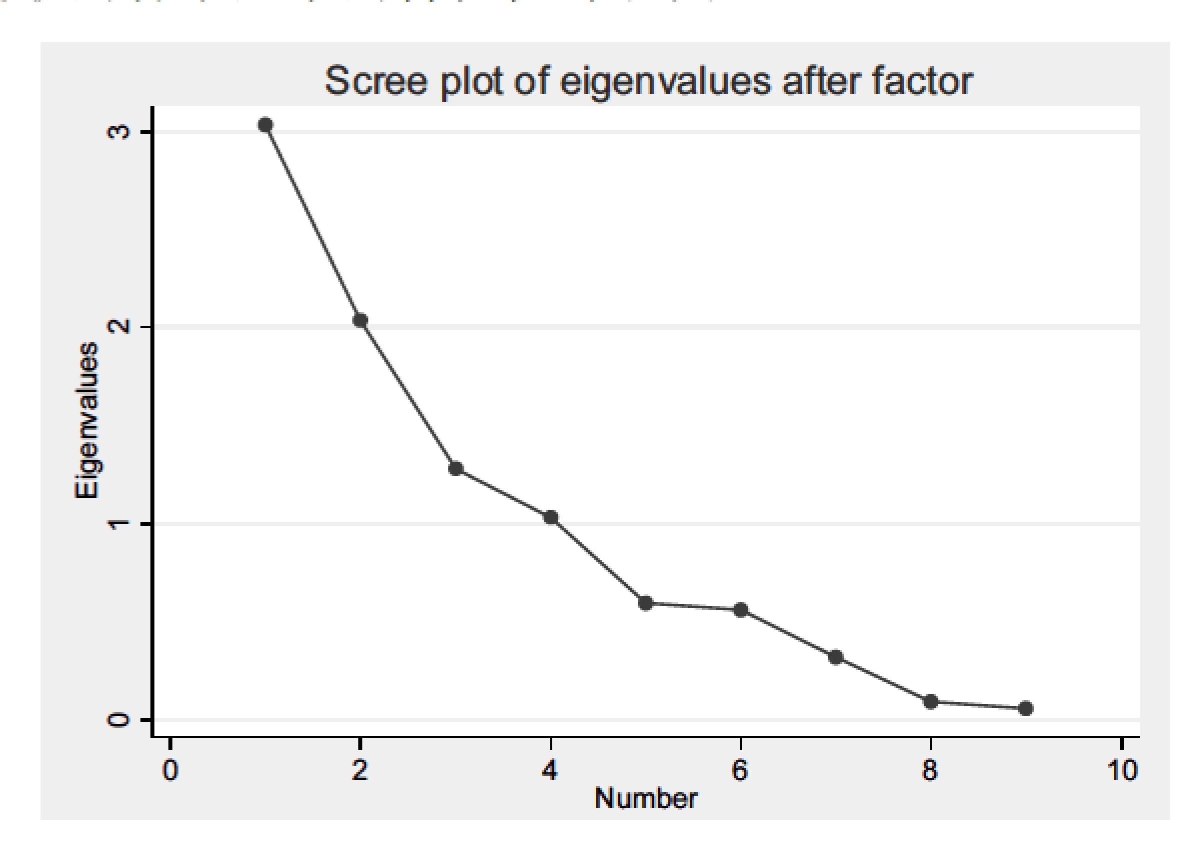


图 22.53　因子分析结果图 9

碎石图可以非常直观地观测出提取因子的特征值大小情况。图 22.53 的横轴表示的是系统提取因子的名称，并且已经按特征值大小进行降序排列好，纵轴表示因子特征值的大小情况。从图 22.53 中可以轻松地看出本例中只有前 4 个因子的特征值是大于 1 的。

2. 对“谷物单位面积产量”“棉花单位面积产量”“花生单位面积产量”“油菜籽单位面积产量”“芝麻单位面积产量”“黄红麻单位面积产量”“甘蔗单位面积产量”“烤烟单位面积产量”“甜菜单位面积产量”9 种作物单位面积产量提取公因子

操作步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中分别输入如下命令并按键盘上的回车键进行确认。

- factor　var28 var29 var30 var31 var32 var33 var34 var35 var38,pcf：本命令的含义是采用主成分因子法对 9 种作物单位面积产量变量进行因子分析。
- rotate：本命令的含义是采用最大方差正交旋转法对因子结构进行旋转。
- loadingplot,factors(2) yline(0) xline(0)：本命令的含义是绘制因子旋转后的因子载荷图。
- predict f1 f2 f3：本命令的含义是展示因子分析后各个样本的因子得分情况。
- correlate　f1 f2 f3：本命令的含义是展示系统提取的 3 个主因子的相关系数矩阵。
- scoreplot,mlabel(var1) yline(0) xline(0)：本命令的含义是展示每个样本的因子得分示意图。
- estat kmo：本命令的含义是展示本例因子分析的 KMO 检验结果。
- screeplot：本命令的含义是展示本例因子分析所提取的各个因子的特征值碎石图。

03 设置完毕后，等待输出结果。

在 Stata 14.0 主界面的结果窗口可以看到如图 22.54~图 22.62 所示的分析结果。

图 22.54 展示的是因子分析的基本情况。

```
. factor  var28 var29 var30 var31 var32 var33 var34 var35 var38,pcf
(obs=31)

Factor analysis/correlation                    Number of obs    =       31
    Method: principal-component factors        Retained factors =        3
    Rotation: (unrotated)                      Number of params =       24

    --------------------------------------------------------------------------
         Factor  |   Eigenvalue   Difference        Proportion   Cumulative
    -------------+------------------------------------------------------------
        Factor1  |      2.66907      0.71186            0.2966       0.2966
        Factor2  |      1.95722      0.83410            0.2175       0.5140
        Factor3  |      1.12312      0.14748            0.1248       0.6388
        Factor4  |      0.97564      0.28254            0.1084       0.7472
        Factor5  |      0.69310      0.05938            0.0770       0.8242
        Factor6  |      0.63373      0.23696            0.0704       0.8947
        Factor7  |      0.39676      0.05606            0.0441       0.9387
        Factor8  |      0.34070      0.13005            0.0379       0.9766
        Factor9  |      0.21066            .            0.0234       1.0000
    --------------------------------------------------------------------------
    LR test: independent vs. saturated:  chi2(36) =   71.63 Prob>chi2 = 0.0004

Factor loadings (pattern matrix) and unique variances

    -------------------------------------------------------
        Variable |  Factor1   Factor2   Factor3 | Uniqueness
    -------------+------------------------------+-----------
           var28 |   0.6222    0.3808    0.0484 |    0.4655
           var29 |   0.2614    0.6981   -0.2271 |    0.3927
           var30 |   0.7382    0.4135   -0.0782 |    0.2779
           var31 |  -0.0019   -0.0111    0.6844 |    0.5314
           var32 |   0.7760    0.2766    0.1581 |    0.2963
           var33 |   0.7326   -0.3723    0.1796 |    0.2924
           var34 |   0.5862   -0.5687   -0.0169 |    0.3327
           var35 |  -0.1664    0.1734    0.7275 |    0.4129
           var38 |  -0.3980    0.7649    0.0881 |    0.2487
    -------------------------------------------------------
```

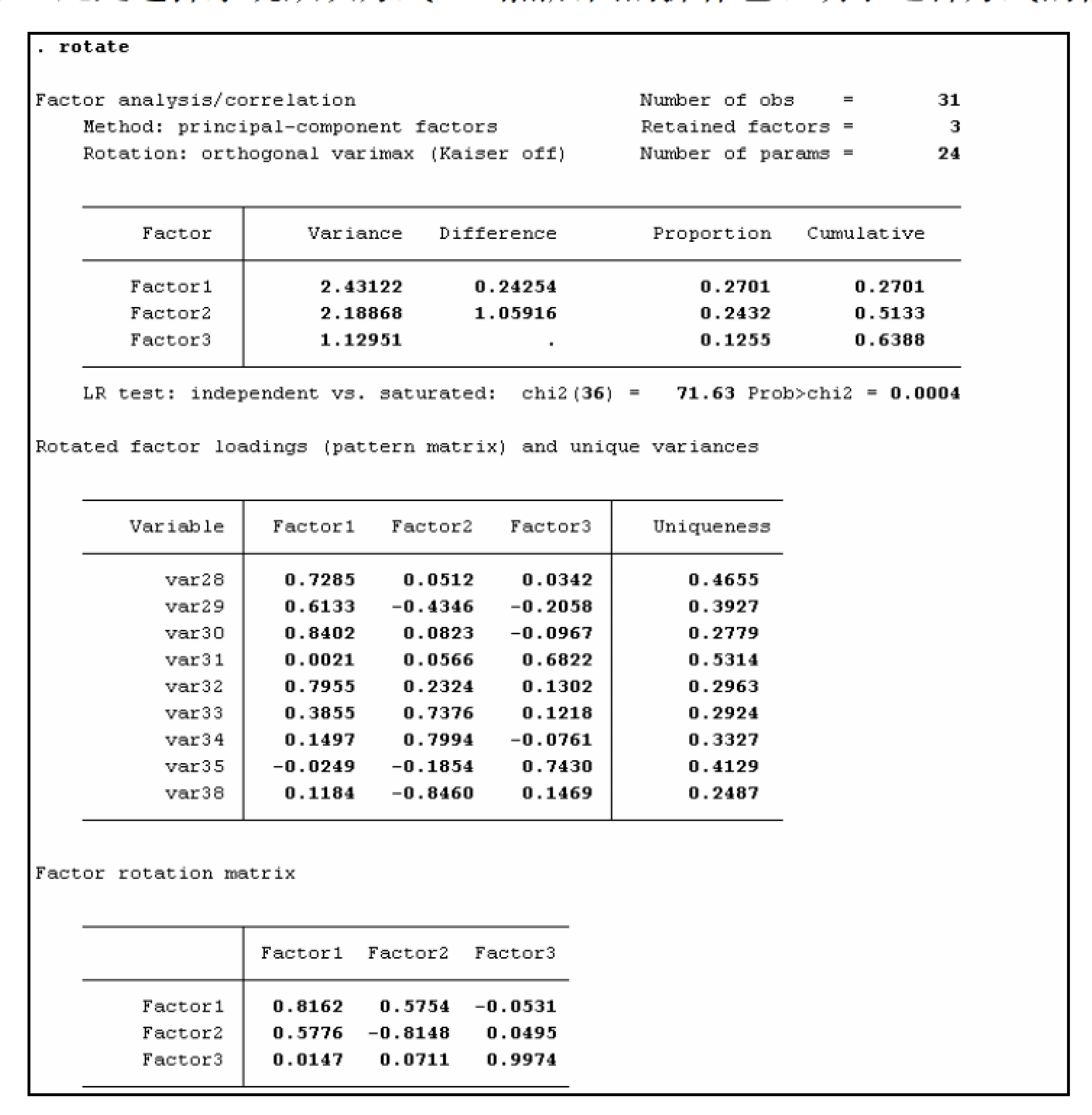
图 22.54　因子分析结果图 10

图 22.54 的上半部分说明的是因子分析模型的一般情况，从图中可以看出共有 31 个样本（Number of obs = 31）参与了分析，提取保留的因子共有 3 个（Retained factors = 3），模型 LR 检验的卡方值（LR test: independent vs. saturated: chi2(36)）为 71.63，P 值（Prob>chi2）为 0.0004，模型非常显著。上半部分最左列（Factor）说明的是因子名称，可以看出模型共提取了 9 个因子。Eigenvalue 列表示提取因子的特征值情况，只有前 3 个因子的特征值是大于 1 的，其中第 1 个因子的特征值是 2.66907，第 2 个因子的特征值是 1.95722。Proportion 列表示的是提取因子的方差贡献率，其中第 1 个因子的方差贡献率为 29.66%，第 2 个因子的方差贡献率为 21.75%。Cumulative 列表示的是提取因子的累计方差贡献率，其中前两个因子的累计方差贡献率为 51.40%。

图 22.54 的下半部分说明的是模型的因子载荷矩阵以及变量的未被解释部分。其中 Variable 列表示的是变量名称，Factor1、Factor2、Factor3 这 3 列分别说明的是提取的前 3 个主因子（特征值大于 1 的）对各个变量的解释程度，本例中，Factor1 主要解释的是 V28、V30、V32、V33、V34 这 5 个变量的信息，Factor2 主要解释的是 V29、V38 变量的信息，Factor3 主要解释的是 V31、V35 这 2 个变量的信息。Uniqueness 列表示变量未被提取的前 4 个主因子解释的部分，可以发现在舍弃其他主因子的情况下，信息的损失量是比较小的。

图 22.55 展示的是对因子结构进行旋转的结果。学者们的研究表明，旋转操作有助于进一步

简化因子结构。Stata 14.0 支持的旋转方式有两种：一种是最大方差正交旋转，一般适用于相互独立的因子或者成分，也是系统默认的情况；另一种是 Promax 斜交旋转，它允许因子或者成分之间存在相关关系。此处选择系统默认方式，当然后面的操作也证明了这种方式的恰当性。



```
. rotate

Factor analysis/correlation                        Number of obs    =        31
    Method: principal-component factors            Retained factors =         3
    Rotation: orthogonal varimax (Kaiser off)       Number of params =        24

    --------------------------------------------------------------------------
         Factor  |     Variance   Difference        Proportion   Cumulative
    -------------+------------------------------------------------------------
        Factor1  |      2.43122      0.24254            0.2701       0.2701
        Factor2  |      2.18868      1.05916            0.2432       0.5133
        Factor3  |      1.12951            .            0.1255       0.6388
    --------------------------------------------------------------------------
    LR test: independent vs. saturated:  chi2(36) =   71.63 Prob>chi2 = 0.0004

Rotated factor loadings (pattern matrix) and unique variances

    -----------------------------------------------------------
        Variable |  Factor1   Factor2   Factor3 |   Uniqueness
    -------------+------------------------------+--------------
           var28 |   0.7285    0.0512    0.0342 |      0.4655
           var29 |   0.6133   -0.4346   -0.2058 |      0.3927
           var30 |   0.8402    0.0823   -0.0967 |      0.2779
           var31 |   0.0021    0.0566    0.6822 |      0.5314
           var32 |   0.7955    0.2324    0.1302 |      0.2963
           var33 |   0.3855    0.7376    0.1218 |      0.2924
           var34 |   0.1497    0.7994   -0.0761 |      0.3327
           var35 |  -0.0249   -0.1854    0.7430 |      0.4129
           var38 |   0.1184   -0.8460    0.1469 |      0.2487
    -----------------------------------------------------------

Factor rotation matrix

    --------------------------------------------
                 |  Factor1  Factor2  Factor3
    -------------+------------------------------
         Factor1 |   0.8162   0.5754  -0.0531
         Factor2 |   0.5776  -0.8148   0.0495
         Factor3 |   0.0147   0.0711   0.9974
    --------------------------------------------
```

图 22.55　因子分析结果图 11

图 22.55 包括 3 部分内容，第 1 部分说明的是因子旋转模型的一般情况，从图中可以看出共有 31 个样本（Number of obs = 31）参与了分析，提取保留的因子共有 3 个（Retained factors = 3），模型 LR 检验的卡方值（LR test: independent vs. saturated: chi2(36)）为 71.63，P 值（Prob>chi2）为 0.0004，模型非常显著。最左列（Factor）说明的是因子名称，可以看出模型旋转后共提取了 3 个因子。Proportion 列表示的是提取因子的方差贡献率，其中第 1 个因子的方差贡献率为 27.01%，第 2 个因子的方差贡献率为 24.32%。Cumulative 列表示的是提取因子的累计方差贡献率，其中前两个因子的累计方差贡献率为 51.33%。

图 22.55 的第 2 部分说明的是模型的因子载荷矩阵以及变量的未被解释部分。其中 Variable 列表示的是变量名称，Factor1、Factor2、Factor3 这 3 列分别说明的是旋转提取的 3 个主因子对各个变量的解释程度，本例中，Factor1 主要解释的是 V28、V29、V30、V32 这 4 个变量的信息，Factor2 主要解释的是 V33、V34、V38 变量的信息，Factor3 主要解释 V31、V35 这 2 个变量的信息。Uniqueness 列表示变量未被提取的前 3 个主因子解释的部分，可以发现在舍弃其他主因子的情况下，信息的损失量是很小的。

图 22.55 的第 3 部分展示的是因子旋转矩阵的一般情况，提取的 3 个因子相关关系不明显。

图 22.56 展示的是因子旋转后的因子载荷图。因子载荷图可以使用户更加直观地看出各个变量被前两个因子解释的情况。

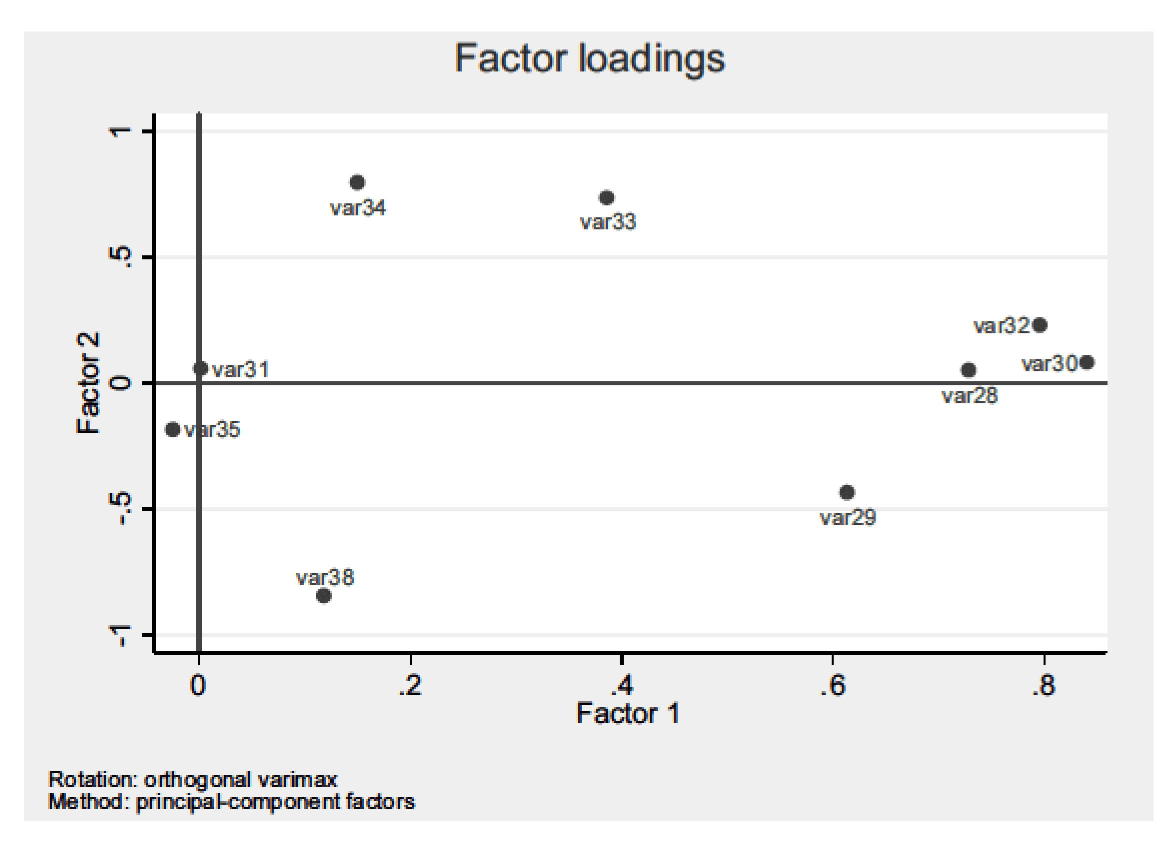


图 22.56　因子分析结果图 12

与前面的分析相同， Factor1 主要解释的是 V28、V29、V30、V32 这 4 个变量的信息，Factor2 主要解释的是 V33、V34、V38 变量的信息。

图 22.57 展示的是因子分析后各个样本的因子得分情况。因子得分的概念是通过将每个变量标准化为平均数等于 0 和方差等于 1，然后以因子分析系数进行加权合计为每个因子构成的线性情况。以因子的方差贡献率为权数对因子进行加权求和，即可得到每个样本的因子综合得分。

```
. predict f1 f2 f3
(regression scoring assumed)

Scoring coefficients (method = regression; based on varimax rotated factors)
```

Variable	Factor1	Factor2	Factor3
var28	0.30328	-0.02130	0.04028
var29	0.28300	-0.24865	-0.18922
var30	0.34675	-0.01794	-0.07367
var31	0.00511	0.04752	0.60756
var32	0.32100	0.06214	0.13201
var33	0.11651	0.32427	0.13553
var34	0.01118	0.36204	-0.04104
var35	0.00985	-0.06204	0.65376
var38	0.10523	-0.39866	0.10552

图 22.57　因子分析结果图 13

根据图 22.57 展示的因子得分系数矩阵，可以写出各公因子的表达式。值得一提的是，在表达式中各个变量已经不是原始变量而是标准化变量。

表达式如下：

F1= 0.30328*谷物单位面积产量+ 0.28300*棉花单位面积产量+ 0.34675*花生单位面积产量+ 0.00511*油菜籽单位面积产量+ 0.32100*芝麻单位面积产量+ 0.11651*黄红麻单位面积产量+ 0.01118*甘蔗单位面积产量+ 0.00985*烤烟单位面积产量+ 0.10523*甜菜单位面积产量

F2= -0.02130*谷物单位面积产量-0.24865*棉花单位面积产量-0.01794*花生单位面积产量 + 0.04752 *油菜籽单位面积产量+0.06214*芝麻单位面积产量+0.32427*黄红麻单位面积产量+0.36204*甘蔗单位面积产量-0.06204*烤烟单位面积产量-0.39866*甜菜单位面积产量

F3= 0.04028*谷物单位面积产量-0.18922*棉花单位面积产量-0.07367　*花生单位面积产量+0.60756*油菜籽单位面积产量+0.13201*芝麻单位面积产量+0.13553*黄红麻单位面积产量-0.04104*甘蔗单位面积产量+0.65376*烤烟单位面积产量+ 0.10552*甜菜单位面积产量

选择“Data” | “Data Editor” | “Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 22.58 所示的因子得分数据。

图 22.59 展示的是系统提取的 3 个主因子的相关系数矩阵。

	var33	var34	var35	var36	var37	var38	f1	f2	f3
1	0	0	0	56.1	18.4	0	-.0114508	-.4903337	-2.352606
2	0	0	0	8.1	3.2	0	.3351588	-.4508	-2.418312
3	2219.4	0	1816.2	1383.3	528.1	37242.9	.5209238	-.7796363	.0181677
4	0	0	3428.4	1015	544.8	57815.9	-.3785098	-1.692973	.332224
5	0	0	4120	2036.6	908.7	40221.8	-.5729568	-1.557982	.5521588
6	0	0	2733.3	450.2	173.1	44482	1.40301	-1.537476	.8721575
7	0	0	2745.1	616.4	221.7	32481.8	1.276931	-1.513297	-.9304208
8	0	0	2440.4	1536.8	682.5	33526.3	-.1466161	-.5833524	1.514766
9	0	64082.4	0	24.3	9.2	0	.4739915	.0949107	-1.039694
10	4600	58940.5	1700	1032.1	332.7	0	1.083796	1.107141	.706693
11	3596.4	62764.4	0	431.1	159.3	0	1.040084	.8796219	-.6305587
12	2997.9	39924.7	2701.8	1317.2	198.5	0	.7144414	.4972804	.5542756
13	3275.4	60904.1	2110.9	133.1	48.3	0	.0505328	.9622147	.0665452
14	4490.9	44891.1	2294.8	1075.3	426.7	0	.4568957	.4706759	-.2653826
15	6700	0	2622.5	2117.2	415.6	0	1.323404	.6571792	1.457165
16	5352.8	67398.2	2345	1477.6	380.2	0	.8995256	1.261619	.6473755
17	3430	41651.7	2024.3	2580	790.3	0	.8932484	.6601623	.4339671
18	3096.6	49813.7	2353.3	2374.8	955	0	.5173347	.6201198	.2538124
19	2365.6	86735.8	2301.9	501.7	118.3	0	-.4300438	1.404414	.0931077
20	2570.9	66599.2	1734.2	1437.9	638.1	0	-.1537787	.8261811	-.5793152
21	6846.7	64072.7	714.3	516.6	198	0	-.7223284	1.800673	-1.424854
22	1390.4	34902.7	1951.3	816	280.5	0	-.4155699	.3922764	.2352655
23	2216.2	46945	2037.4	1528.2	720.7	17020.4	.0998291	.2783942	.702662
24	666.7	36409.9	1625.1	2570.2	1363.7	4365.9	-1.376114	.2806397	-.2792853

图 22.58　因子分析结果图 14

```
. correlate  f1 f2 f3
(obs=31)

             |      f1       f2       f3
-------------+---------------------------
          f1 |  1.0000
          f2 |  0.0000   1.0000
          f3 | -0.0000  -0.0000   1.0000
```

图 22.59　因子分析结果图 15

从图 22.59 中可以看出，提取的 3 个主因子之间几乎没有什么相关关系，这也说明了在前面对因子进行旋转的操作环节中采用最大方差正交旋转方式是明智的。值得说明的是，图中有的相关系数是-0.0000 并非是不正确的，这是因为 Stata 14.0 只保留了 4 位小数所致，例如真实的数据有可能是-0.00001，那么结果显示的就是-0.0000。

图 22.60 展示的是每个样本在前两个主因子维度上的因子得分示意图。

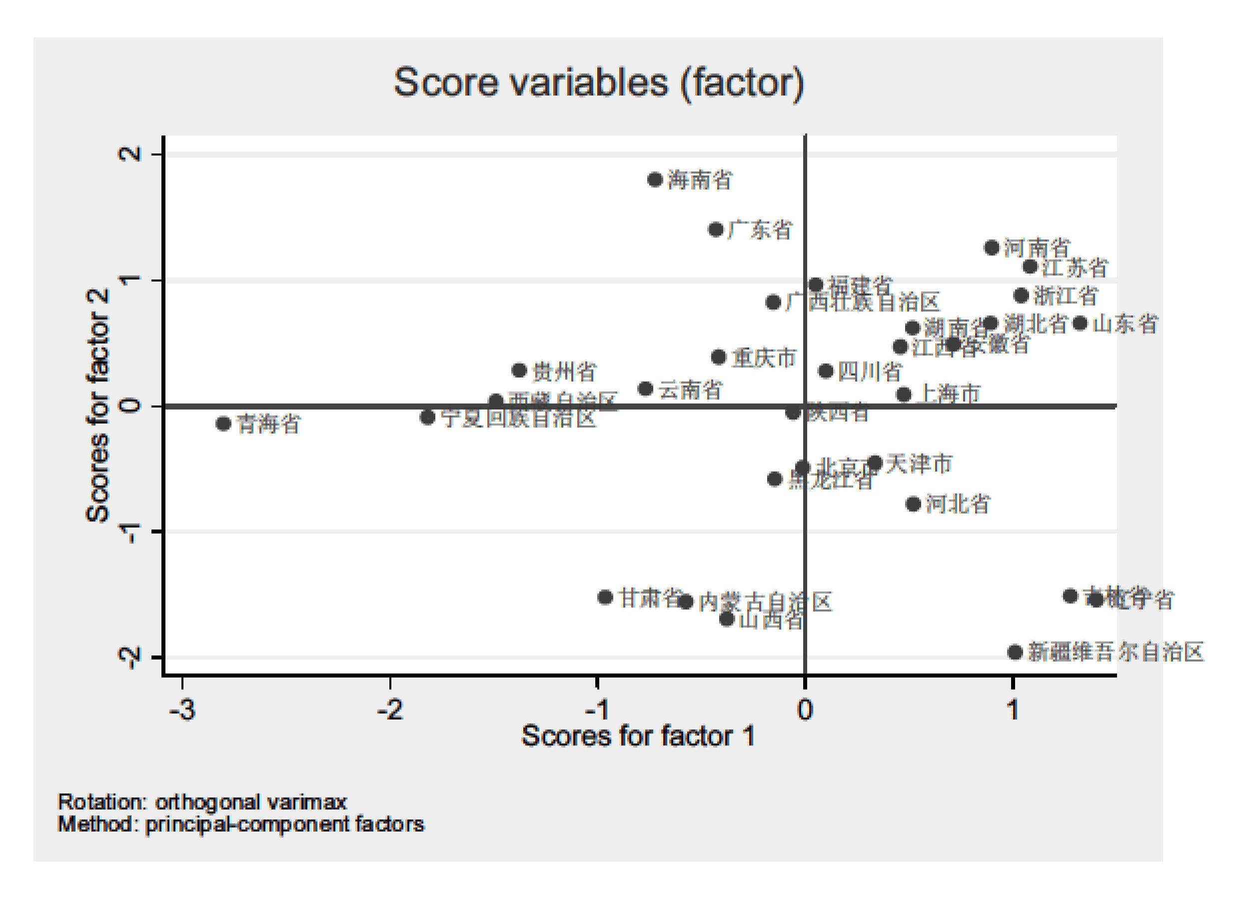


图 22.60　因子分析结果图 16

从图 22.60 中可以看出，所有的样本被分到 4 个象限，可以比较直观地看出各个样本的因

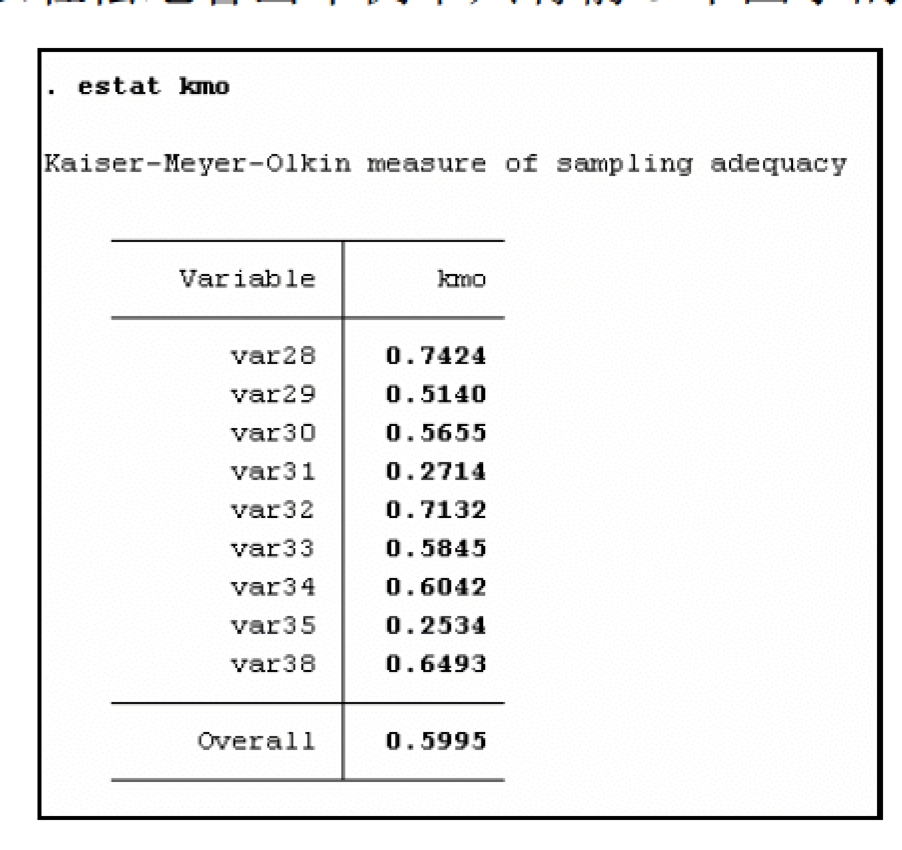
子得分分布情况。

图 22.61 展示的是本例因子分析的 KMO 检验结果。

KMO 检验是为了查看数据是否适合进行因子分析，其取值范围是 0~1。其中，0.9~1 表示极好，0.8~0.9 表示可奖励的，0.7~0.8 表示还好，0.6~0.7 表示中等。本例中总体（Overall）KMO 的取值为 0.5995，表明因子分析的效果是差强人意的。

图 22.62 展示的是本例因子分析所提取的各个因子的特征值碎石图。

碎石图可以非常直观地观测出提取因子的特征值的大小情况。横轴表示的是系统提取因子的名称，并且已经按特征值大小进行降序排列，纵轴表示因子特征值的大小情况。从图中可以轻松地看出本例中只有前 3 个因子的特征值是大于 1 的。



```
. estat kmo

Kaiser-Meyer-Olkin measure of sampling adequacy
```

Variable	kmo
var28	0.7424
var29	0.5140
var30	0.5655
var31	0.2714
var32	0.7132
var33	0.5845
var34	0.6042
var35	0.2534
var38	0.6493
Overall	0.5995

图 22.61　因子分析结果图 17

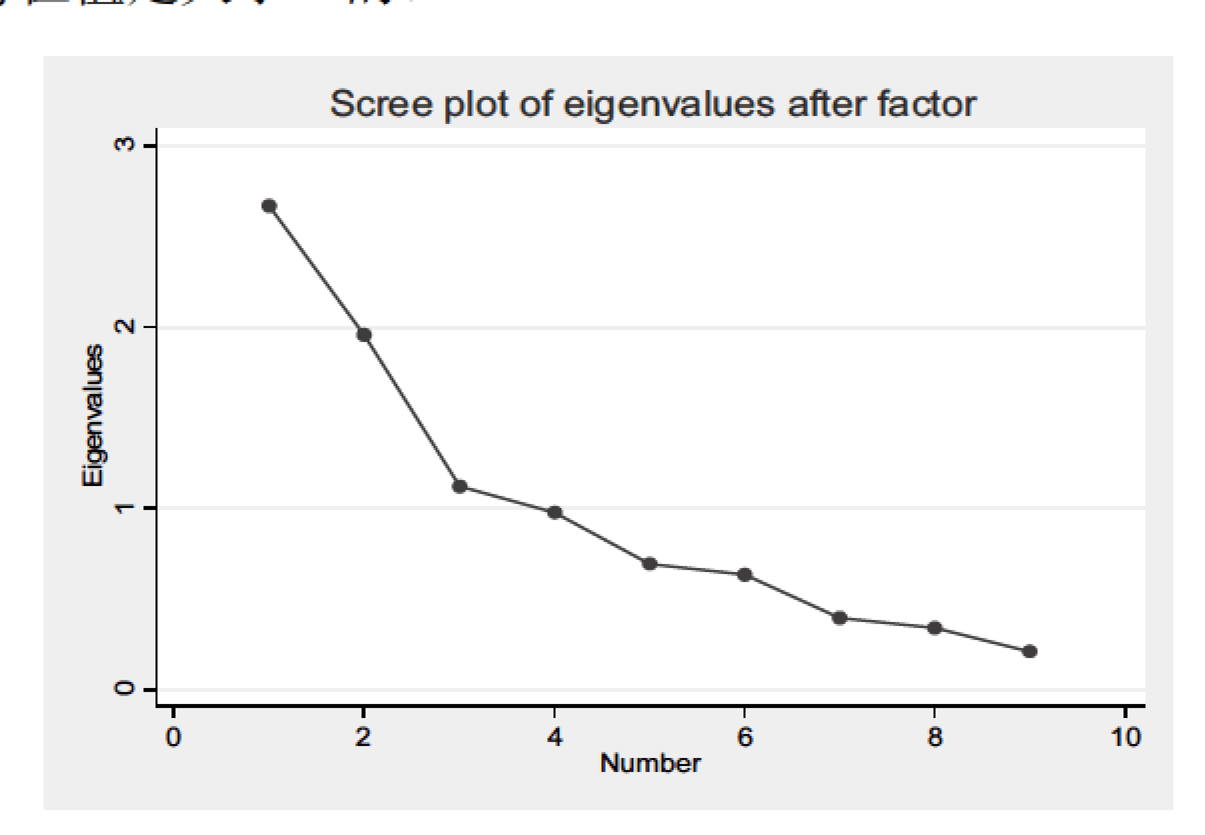


图 22.62　因子分析结果图 18

22.8　聚类分析

对于聚类分析，准备从 3 部分进行操作：

- 对粮食产品的组成部分（包括“稻谷”“小麦”“玉米”“豆类”“薯类”）变量进行聚类。
- 对水果产品的组成部分（包括“苹果”“柑桔”“梨”“葡萄”“香蕉”）变量进行聚类。
- 对油料作物的组成部分（包括“花生”“油菜籽”“芝麻”）变量进行聚类。

1. 对粮食产品的组成部分包括“稻谷”“小麦”“玉米”“豆类”“薯类”等变量进行聚类

观察到不同变量的数量级相差不大，所以无须先对数据进行标准化处理，直接进行分析即可。

分析步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中分别输入如下命令并按键盘上的回车键进行确认。本操作命令的含义是设定聚类数为 3，然后使用“K 个平均数的聚类分析”方法对粮食产品的组成部分（包括“稻谷”“小麦”“玉米”“豆类”“薯类”）变量进行分析。

```
cluster kmeans var13 var14 var15 var16 var17,k(3)
```

03 设置完毕后，按键盘上的回车键，等待输出结果。

在 Stata 14.0 主界面的结果窗口可以看到如图 22.63~图 22.65 所示的分析结果。

图 22.63 展示的是设定聚类数为 3，然后使用“K 个平均数的聚类分析”方法进行分析的结果。在输入 Stata 命令并且分别按键盘上的回车键进行确认后，可以看到系统产生了一个新的变量：聚类变量_clus_1（cluster name: _clus_1）。

```
. cluster kmeans var13 var14 var15 var16 var17,k(3)
cluster name: _clus_1
```

图 22.63　聚类分析结果图 1

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 22.64 所示的聚类数据。

	var1	var13	var14	var15	var16	var17	_clus_1
1	北京市	.2	28.4	90.3	1.2	1.3	2
2	天津市	10.7	54.2	94.4	1.7	.6	2
3	河北省	60.2	1276.1	1639.6	35.7	104.6	1
4	山西省	.5	240.3	854.6	24.4	30.2	2
5	内蒙古自治区	77.9	170.9	1632.1	171.3	204	2
6	辽宁省	505.1	3.7	1360.3	37	65	2
7	吉林省	623.5	1.3	2339	101.3	54.5	3
8	黑龙江省	2062.1	103.8	2675.8	577.8	134.7	3
9	上海市	88.9	24.1	2.8	1.5	.8	2
10	江苏省	1864.2	1023.2	226.2	82.8	38.6	1
11	浙江省	649	27	14.6	31.6	45.4	2
12	安徽省	1387.1	1215.7	362.6	115	46.5	1
13	福建省	514.1	.8	16.6	19.9	119.7	2
14	江西省	1950.1	2.2	10.5	28.7	59.9	1
15	山东省	104	2103.9	1978.7	43.3	188.1	1
16	河南省	474.5	3123	1696.5	95.2	139.3	1
17	湖北省	1616.9	344.8	276.2	39.5	99.8	1
18	湖南省	2575.4	10.2	188.5	41.1	118.8	1
19	广东省	1096.9	.3	78.9	18.2	164.2	2
20	广西壮族自治区	1084.1	.2	244.7	28.9	67.8	2
21	海南省	145.1	0	10.3	2.3	30.3	2
22	重庆市	493.5	42.4	257	43.5	284.2	2
23	四川省	1527.1	436	701.6	96.2	441.7	1
24	贵州省	303.9	50.4	243.7	22	239.3	2
25	云南省	668.7	98.9	598.2	125.7	178.9	2
26	西藏自治区	.6	24.9	2.8	2.4	.4	2
27	陕西省	84.5	410.9	550.7	44.6	80.3	2
28	甘肃省	0	247.5	425.6	34.8	228.9	2
29	青海省	0	35.4	15.2	7.1	36.9	2
30	宁夏回族自治区	70.8	63	172.4	4.7	44.5	2
31	新疆维吾尔自治区	60.6	576.6	517.7	29.1	24	2

图 22.64　聚类分析结果图 2

从图 22.64 中可以看到所有的观测样本被分为 3 类。其中，江西、河南、山东、江苏、四川、湖南、河北、湖北、安徽被分到第 1 类，吉林、黑龙江被分到第 3 类，其他的省市被分到第 2 类。

为观测 3 类样本的特征，可以对数据进行排序操作，在主界面的“Command”文本框中输入操作命令：

```
sort _clus_1
```

并按键盘上的回车键进行确认，然后选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 22.65 所示整理后的数据。

	var1	var13	var14	var15	var16	var17	_clus_1
1	江西省	1950.1	2.2	10.5	28.7	59.9	1
2	河南省	474.5	3123	1696.5	95.2	139.3	1
3	山东省	104	2103.9	1978.7	43.3	188.1	1
4	江苏省	1864.2	1023.2	226.2	82.8	38.6	1
5	四川省	1527.1	436	701.6	96.2	441.7	1
6	湖南省	2575.4	10.2	188.5	41.1	118.8	1
7	河北省	60.2	1276.1	1639.6	35.7	104.6	1
8	湖北省	1616.9	344.8	276.2	39.5	99.8	1
9	安徽省	1387.1	1215.7	362.6	115	46.5	1
10	内蒙古自治区	77.9	170.9	1632.1	171.3	204	2
11	山西省	.5	240.3	854.6	24.4	30.2	2
12	广东省	1096.9	.3	78.9	18.2	164.2	2
13	辽宁省	505.1	3.7	1360.3	37	65	2
14	重庆市	493.5	42.4	257	43.5	284.2	2
15	云南省	668.7	98.9	598.2	125.7	178.9	2
16	广西壮族自治区	1084.1	.2	244.7	28.9	67.8	2
17	北京市	.2	28.4	90.3	1.2	1.3	2
18	青海省	0	35.4	15.2	7.1	36.9	2
19	海南省	145.1	0	10.3	2.3	30.3	2
20	宁夏回族自治区	70.8	63	172.4	4.7	44.5	2
21	贵州省	303.9	50.4	243.7	22	239.3	2
22	陕西省	84.5	410.9	550.7	44.6	80.3	2
23	福建省	514.1	.8	16.6	19.9	119.7	2
24	新疆维吾尔自治区	60.6	576.6	517.7	29.1	24	2
25	甘肃省	0	247.5	425.6	34.8	228.9	2
26	西藏自治区	.6	24.9	2.8	2.4	.4	2
27	天津市	10.7	54.2	94.4	1.7	.6	2
28	浙江省	649	27	14.6	31.6	45.4	2
29	上海市	88.9	24.1	2.8	1.5	.8	2
30	吉林省	623.5	1.3	2339	101.3	54.5	3
31	黑龙江省	2062.1	103.8	2675.8	577.8	134.7	3

图 22.65　聚类分析结果图 3

可以看到第 1 类样本的特征是各类粮食作物的产量普遍较高，第 3 类样本的特征是稻谷、玉米、豆类的产量大多比较高，第 2 类样本的特征不明显，但综合来看各种作物的产量都比较低。

2. 对水果产品的组成部分（包括“苹果”“柑桔”“梨”“葡萄”“香蕉”）变量进行聚类

观察到不同变量的数量级相差不大，所以无须先对数据进行标准化处理，直接进行分析即可。

分析步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中分别输入如下命令并按键盘上的回车键进行确认。本操作命令的含义是设定聚类数为 3，然后使用“K 个平均数的聚类分析”方法对“苹果”“柑桔”“梨”“葡萄”“香蕉”等变量进行分析。

```
cluster kmeans  var23 var24 var25 var26 var27,k(3)
```

03 设置完毕后，按键盘上的回车键，等待输出结果。

在 Stata 14.0 主界面的结果窗口可以看到如图 22.66~图 22.68 所示的分析结果。

图 22.66 展示的是设定聚类数为 3，然后使用“K 个平均数的聚类分析”方法进行分析的结果。在输入 Stata 命令并且分别按键盘上的回车键进行确认后，可以看到系统产生了一个新的变量：聚类变量_clus_2（cluster name: _clus_2）。

```
. cluster kmeans  var23 var24 var25 var26 var27,k(3)
cluster name: _clus_2
```

图 22.66　聚类分析结果图 4

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 22.67 所示的_clus_2 数据。

	var1	var23	var24	var25	var26	var27	_clus_2
1	江西省	0	356.7	13.5	3.3	0	1
2	河南省	420.3	3.9	100.5	50.1	0	2
3	山东省	837.9	0	122.7	98.5	0	3
4	江苏省	61.7	5.1	73	39.2	0	1
5	四川省	45.7	319.4	92.3	24.3	3.6	1
6	湖南省	0	420.4	15.1	11.9	0	1
7	河北省	292.6	0	406.9	112.5	0	2
8	湖北省	1	331	46.3	15.2	0	1
9	安徽省	41.1	2.9	100.4	25.9	0	1
10	内蒙古自治区	10.6	0	7.7	7.4	0	1
11	山西省	333.9	0	59	25.9	0	2
12	广东省	0	378.7	7.4	0	384.9	1
13	辽宁省	239.7	0	140.2	67.3	0	2
14	重庆市	.6	153.3	30.4	5.4	.2	1
15	云南省	25.3	45	36.4	35.6	168.7	1
16	广西壮族自治区	0	355	24.2	27.2	205.7	1
17	北京市	10.5	0	16.2	4.2	0	1
18	青海省	.6	0	0	0	0	1
19	海南省	0	4.5	0	0	189.2	1
20	宁夏回族自治区	40.9	0	2.9	14.1	0	1
21	贵州省	2.2	20.8	19.5	8	.6	1
22	陕西省	902.9	34.3	88.1	36.4	0	3
23	福建省	0	300.4	19.7	11.2	87	1
24	新疆维吾尔自治区	71.5	0	60.6	175.5	0	1
25	甘肃省	227.6	.4	33.4	12.5	0	2
26	西藏自治区	.5	0	.1	0	0	1
27	天津市	5.5	0	3.9	12.3	0	1
28	浙江省	0	194.4	38.6	52.7	0	1

图 22.67　聚类分析结果图 5

从图 22.67 中可以看到所有的观测样本被分为 3 类。其中，河南、辽宁、陕西、甘肃、河北被分到第 2 类，山东、陕西被分到第 3 类，其他的省市被分到第 1 类。

为观测 3 类样本的特征，可以对数据进行排序操作，在主界面的“Command”文本框中输入操作命令：

```
sort _clus_2
```

并按键盘上的回车键进行确认，然后选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 22.68 所示整理后的数据。

	var1	var23	var24	var25	var26	var27	_clus_2
1	北京市	10.5	0	16.2	4.2	0	1
2	重庆市	.6	153.3	30.4	5.4	.2	1
3	贵州省	2.2	20.8	19.5	8	.6	1
4	宁夏回族自治区	40.9	0	2.9	14.1	0	1
5	海南省	0	4.5	0	0	189.2	1
6	四川省	45.7	319.4	92.3	24.3	3.6	1
7	安徽省	41.1	2.9	100.4	25.9	0	1
8	湖南省	0	420.4	15.1	11.9	0	1
9	青海省	.6	0	0	0	0	1
10	广西壮族自治区	0	355	24.2	27.2	205.7	1
11	西藏自治区	.5	0	.1	0	0	1
12	江西省	0	356.7	13.5	3.3	0	1
13	云南省	25.3	45	36.4	35.6	168.7	1
14	浙江省	0	194.4	38.6	52.7	0	1
15	内蒙古自治区	10.6	0	7.7	7.4	0	1
16	新疆维吾尔自治区	71.5	0	60.6	175.5	0	1
17	江苏省	61.7	5.1	73	39.2	0	1
18	福建省	0	300.4	19.7	11.2	87	1
19	上海市	0	17.7	3.2	9.5	0	1
20	湖北省	1	331	46.3	15.2	0	1
21	吉林省	14.4	0	13.3	14.2	0	1
22	黑龙江省	11.4	0	4	6.2	0	1
23	广东省	0	378.7	7.4	0	384.9	1
24	天津市	5.5	0	3.9	12.3	0	1
25	河南省	420.3	3.9	100.5	50.1	0	2
26	辽宁省	239.7	0	140.2	67.3	0	2
27	山西省	333.9	0	59	25.9	0	2
28	甘肃省	227.6	.4	33.4	12.5	0	2
29	河北省	292.6	0	406.9	112.5	0	2
30	山东省	837.9	0	122.7	98.5	0	3
31	陕西省	902.9	34.3	88.1	36.4	0	3

图 22.68　聚类分析结果图 6

可以看到第 1 类样本的特征是各种水果产品的产量比较低，第 3 类样本的特征是苹果的产量非常高，葡萄和梨的产量比较高，第 2 类样本的特征是各种作物的产量都比较高。

3. 对油料作物的组成部分（包括“花生”“油菜籽”“芝麻”）变量进行聚类

观察到不同变量的数量级相差不大，所以无须先对数据进行标准化处理，直接进行分析即可。分析步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出主界面。

02 在主界面的“Command”文本框中分别输入如下命令并按键盘上的回车键进行确认，本操作命令的含义是设定聚类数为 3，然后使用“K 个平均数的聚类分析”方法对“花生”“油菜籽”“芝麻”等变量进行分析。

```
cluster kmeans   var18 var19 var20,k(3)
```

03 设置完毕后，按键盘上的回车键，等待输出结果。

在 Stata 14.0 主界面的结果窗口可以看到如图 22.69~图 22.71 所示的分析结果。

图 22.69 展示的是设定聚类数为 3，然后使用“K 个平均数的聚类分析”方法进行分析的结果。在输入 Stata 命令并且分别按键盘上的回车键进行确认后，可以看到系统产生了一个新的变量：聚类变量_clus_3（cluster name: _clus_3）。

```
. cluster kmeans   var18 var19 var20,k(3)
cluster name: _clus_3
```

图 22.69　聚类分析结果图 7

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 22.70 所示的_clus_3 数据。

	var1	var18	var19	var20	_clus_3
2	重庆市	10.1	35.1	.7	1
3	贵州省	6.1	71.8	0	1
4	宁夏回族自治区	0	.1	0	1
5	海南省	9.8	0	.2	1
6	四川省	62.7	214.4	.5	3
7	安徽省	84.3	122.8	6.2	3
8	湖南省	32	182	1.4	3
9	青海省	0	32.7	0	1
10	广西壮族自治区	47.5	1.6	.6	1
11	西藏自治区	0	6.3	0	1
12	江西省	43.7	66.7	3.2	1
13	云南省	7	51.8	0	1
14	浙江省	5.4	33.6	.9	1
15	内蒙古自治区	3.1	24	.2	1
16	新疆维吾尔自治区	1.3	15.2	.1	1
17	江苏省	37	105.2	1.8	3
18	福建省	25.7	1.6	.2	1
19	上海市	.2	1.6	0	1
20	湖北省	68.7	220.4	14.6	3
21	吉林省	36	0	1.4	1
22	黑龙江省	5.7	.1	.1	1
23	广东省	90.8	.8	.3	1
24	天津市	.5	0	0	1
25	河南省	429.8	77.3	24.1	2
26	辽宁省	116.5	.1	.2	1
27	山西省	2.2	.6	.5	1
28	甘肃省	.3	33.1	0	1
29	河北省	128.9	3	1	1
30	山东省	338.6	2.2	.1	2
31	陕西省	9.3	38.4	2.4	1

图 22.70　聚类分析结果图 8

从图 22.70 中可以看到所有的观测样本被分为 3 类。其中，河南、山东被分到第 2 类，安徽、四川、湖北、湖南、江苏被分到第 3 类，其他的省市被分到第 1 类。

为观测 3 类样本的特征，可以对数据进行排序操作，在主界面的“Command”文本框中输入操作命令：

```
sort _clus_3
```

并按键盘上的回车键进行确认，然后选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 22.71 所示的整理后的数据。

	var1	var18	var19	var20	_clus_3
2	新疆维吾尔自治区	1.3	15.2	.1	1
3	广西壮族自治区	47.5	1.6	.6	1
4	天津市	.5	0	0	1
5	黑龙江省	5.7	.1	.1	1
6	内蒙古自治区	3.1	24	.2	1
7	福建省	25.7	1.6	.2	1
8	上海市	.2	1.6	0	1
9	青海省	0	32.7	0	1
10	山西省	2.2	.6	.5	1
11	浙江省	5.4	33.6	.9	1
12	贵州省	6.1	71.8	0	1
13	西藏自治区	0	6.3	0	1
14	宁夏回族自治区	0	.1	0	1
15	吉林省	36	0	1.4	1
16	陕西省	9.3	38.4	2.4	1
17	云南省	7	51.8	0	1
18	广东省	90.8	.8	.3	1
19	重庆市	10.1	35.1	.7	1
20	河北省	128.9	3	1	1
21	江西省	43.7	66.7	3.2	1
22	甘肃省	.3	33.1	0	1
23	北京市	1.3	0	0	1
24	海南省	9.8	0	.2	1
25	河南省	429.8	77.3	24.1	2
26	山东省	338.6	2.2	.1	2
27	安徽省	84.3	122.8	6.2	3
28	四川省	62.7	214.4	.5	3
29	湖北省	68.7	220.4	14.6	3

图 22.71　聚类分析结果图 9

可以看到第 3 类样本的特征是油菜籽的产量非常高，花生和芝麻的产量比较高，第 2 类样本的特征是花生的产量非常高，第 1 类样本的特征是各种作物的产量比较低。

通过聚类分析得到的研究结论如下。

- 江西、河南、山东、江苏、四川、湖南、河北、湖北、安徽等省市各类粮食作物的产量普遍较高，吉林、黑龙江等省市稻谷、玉米、豆类的产量大都比较高，其他的省市综合来看各种粮食作物的产量比较低。
- 山东、陕西等省市苹果的产量非常高、葡萄和梨的产量比较高，河南、辽宁、陕西、甘肃、河北等省市各种作物的产量比较高，其他的省市各种水果产品的产量比较低。
- 安徽、四川、湖北、湖南、江苏等省市油菜籽的产量非常高，花生和芝麻的产量比较高，河南、山东等省市花生的产量非常高，其他的省市综合来看各种油料作物的产量比较低。

22.9 研究结论

根据以上所做的分析，可以比较有把握地得出以下结论：

- 简单相关分析表明："农业总产值"的 9 个来源中，"粮食产量"与"油料产量"、"棉花产量"与"甜菜产量"、"油料产量"与"麻类产量"等变量之间的相关性在 1%的显著性水平上显著。
- 简单相关分析表明：9 种农产品的单位面积产量等变量之间的相关性都比较差，都在 0.01 的显著性水平上不显著。
- 简单相关分析表明："稻谷""小麦""玉米""豆类""薯类"5 种粮食作物中仅有"玉米"与"豆类"之间的相关性在 1%的显著性水平上显著。
- 简单相关分析表明："花生""油菜籽""芝麻"3 种油料作物中，仅有"花生"与"芝麻"之间的相关性在 1%的显著性水平上显著。
- 简单相关分析表明："苹果""柑桔""梨""葡萄""香蕉"5 种水果产品中，仅有"梨"与"葡萄"变量之间的相关性在 1%的显著性水平上显著。
- 经过多重线性回归分析，可以发现我国农业总产值水平与"粮食产量""棉花产量""甜菜产量""茶叶产量"以及"水果产量"都有一定的显著关系。具体而言，"粮食产量""棉花产量""茶叶产量"以及"水果产量"有拉动效应，尤其是茶叶产量，每增加一个单位会带来对应农业总产值的 18 倍多的增加；甜菜产量对农业总产值水平有拖后效应，在一定程度上说明种植这种作物是不经济的。
- 经过多重线性回归分析，可以发现我国农业总产值水平与"花生单位面积产量""油菜籽单位面积产量""芝麻单位面积产量""黄红麻单位面积产量"以及"受灾面积(千公顷)"都有一定的显著关系。具体而言，这些变量都对我国的农业总产值有显著拉动效应。"花生单位面积产量""油菜籽单位面积产量""芝麻单位面积产量""黄红麻单位面积产量"对我国的农业总产值有显著拉动效应说明这些作物都是经济的，也就是量的提高能够带来价值的提高，"受灾面积(千公顷)"对我国的农业总产值有显著拉动效应说明"谷贱伤农"的道理在我国是存在的，受灾面积的扩大会带来产量的降低，但这却能带来价格的提高，而且价格提高的幅度要更大，造成总价值也会提高。
- 因子分析表明：可以对"粮食产量""棉花产量""油料产量""麻类产量""甘蔗产量"、"甜菜产量""烟叶产量""茶叶产量""水果产量"9 种农产品产量变量提取 4 个公因子。
- 因子分析表明：可以对"谷物单位面积产量""棉花单位面积产量""花生单位面积产量""油菜籽单位面积产量""芝麻单位面积产量""黄红麻单位面积产量""甘蔗单位面积产量""烤烟单位面积产量""甜菜单位面积产量"9 种作物单位面积产量提取 3 个公因子。
- 聚类分析表明：江西、河南、山东、江苏、四川、湖南、河北、湖北、安徽等省市各类粮食作物的产量普遍较高，吉林、黑龙江等省市稻谷、玉米、豆类的产量大都比较高，其他的省市综合来看各种粮食作物的产量比较低。
- 聚类分析表明：山东、陕西等省市苹果的产量非常高，葡萄和梨的产量比较高，河南、

辽宁、陕西、甘肃、河北等省市的各种水果产品的产量比较高，其他的省市各种水果产品的产量比较低。

- 聚类分析表明：安徽、四川、湖北、湖南、江苏等省市油菜籽的产量非常高，花生和芝麻的产量比较高，河南、山东等省市花生的产量非常高，综合来看其他的省市各种油料作物的产量比较低。

经过以上研究，可以从一种宏观的视野上对我国的农业有一个比较全面的了解，这对于以后我国农业的发展有重要的借鉴和指导意义。例如根据回归分析部分的结论，“受灾面积（千公顷）”对我国的农业总产值有显著拉动效应，说明“谷贱伤农”的道理在我国是存在的，所以继续需要付出更多的努力来保障农业劳动者的利益。再如，聚类分析表明，山东、陕西等省市苹果的产量非常高，葡萄和梨的产量比较高，水果销售商可以据此制定自己的采购渠道建设和物流运输计划。

22.10 本章习题

使用《中国统计年鉴 2011》上的中国各省市 2010 年农产品的相关数据，包括“农业总产值”“粮食产量”“棉花产量”“油料产量”“麻类产量”“甘蔗产量”“甜菜产量”“烟叶产量”“茶叶产量”“水果产量”“谷物”“稻谷”“小麦”“玉米”“豆类”“薯类”“花生”“油菜籽”“芝麻”“黄红麻”“烤烟”“苹果”“柑桔”“梨”“葡萄”“香蕉”“谷物单位面积产量”“棉花单位面积产量”“花生单位面积产量”“油菜籽单位面积产量”“芝麻单位面积产量”“黄红麻单位面积产量”“甘蔗单位面积产量”“烤烟单位面积产量”“受灾面积（千公顷）”“成灾面积（千公顷）”“甜菜单位面积产量”等（数据已整理入 Stata 中），进行以下分析。

（1）相关分析

- 对“农业总产值”的 9 个来源——“粮食产量”“棉花产量”“油料产量”“麻类产量”“甘蔗产量”“甜菜产量”“烟叶产量”“茶叶产量”“水果产量”进行简单相关分析。
- 对 9 种农产品的单位面积产量——“谷物单位面积产量”“棉花单位面积产量”“花生单位面积产量”“油菜籽单位面积产量”“芝麻单位面积产量”“黄红麻单位面积产量”、“甘蔗单位面积产量”“烤烟单位面积产量”“甜菜单位面积产量”进行简单相关分析。
- 对“稻谷”“小麦”“玉米”“豆类”“薯类”5 种粮食作物进行简单相关分析。
- 对“花生”“油菜籽”“芝麻”3 种油料作物进行简单相关分析。
- 对“苹果”“柑桔”“梨”“葡萄”“香蕉”5 种水果产品进行简单相关分析。

（2）回归分析

- 以“农业总产值”为因变量，以农业为自变量，进行最小二乘线性回归。
- 以“农业总产值”为因变量，以“谷物单位面积产量”“棉花单位面积产量”“花生单位面积产量”“油菜籽单位面积产量”“芝麻单位面积产量”“黄红麻单位面积产量”“甘蔗单位面积产量”“烤烟单位面积产量”“甜菜单位面积产量”“受灾面积（千公顷）”、“成灾面积（千公顷）”为自变量，进行最小二乘线性回归。

（3）因子分析

- 对“粮食产量”“棉花产量”“油料产量”“麻类产量”“甘蔗产量”“甜菜产量”“烟叶产量”“茶叶产量”“水果产量”9种农产品产量变量提取公因子。
- 对“谷物单位面积产量”“棉花单位面积产量”“花生单位面积产量”“油菜籽单位面积产量”“芝麻单位面积产量”“黄红麻单位面积产量”“甘蔗单位面积产量”“烤烟单位面积产量”“甜菜单位面积产量”9种作物单位面积产量提取公因子。

（4）聚类分析

- 对粮食产品的组成部分（包括“稻谷”“小麦”“玉米”“豆类”“薯类”）变量进行聚类。
- 对水果产品的组成部分（包括“苹果”“柑桔”“梨”“葡萄”“香蕉”）变量进行聚类。
- 对油料作物的组成部分（包括“花生”“油菜籽”“芝麻”）变量进行聚类。

第 23 章　Stata 在保险业中的应用

保险是指投保人根据保险合同的约定，向保险人支付保险费，保险人对于合同约定的可能发生的事故因其发生所造成的财产损失承担赔偿责任，或者当被保险人死亡、伤残、疾病或者达到合同约定的年龄、期限时承担给付保险金责任的商业保险行为。保险最基本的功能是经济补偿，有利于受灾企业及时恢复生产，有利于企业加强危险管理，有利于安定人民生活。保险业作为国民经济一个不可或缺的组成部分，在我们建设与完善有中国特色的社会主义市场经济中发挥着越来越重要的作用。Stata 作为一种功能强大的统计分析软件，完全可以用来进行保险业的相关分析研究，定量分析变量之间的联系与区别。下面我们就来介绍一下 Stata 在保险业中的应用。

23.1　研究背景及目的

背景一：进入 21 世纪以来，中国保险业持续快速发展，保险机构个数和保险业从业人数不断增加。

根据《中华人民共和国年鉴 2008》提供的数据（见表 23.1），可以发现，无论是保险机构个数还是保险业从业人数都呈现出持续快速增长趋势。

表 23.1　中国历年保险业机构数和从业人数统计（2000—2007 年）

年份	2000年	2001年	2002年	2003年	2004年	2005年	2006年	2007年
机构数/个	33	35	44	62	68	93	107	120
职工人数/人	166 602	185 502	194 383	199 705	262 429	366 559	434 001	506 223

背景二：伴随着保险机构和从业人员的不断增加，保险业的保费收入也持续增长，使得我国保险业呈现出良好发展的态势。

根据《中华人民共和国年鉴2008》提供的数据（见表23.2），可以发现，不管是财产保险公司还是人寿保险公司的保费收入都不断增长。

表 23.2　中国历年保险业保费收入情况统计（2000—2007 年）

年份	2000年	2001年	2002年	2003年	2004年	2005年	2006年	2007年
保费总收入/亿元	1598	2109	3054	3880	4318	4932	5643	7036
财产保险公司保费收入/亿元	608	685	780	869	1125	1283	1579	2086
人寿保险公司保费收入/亿元	990	1424	2274	3011	3194	3649	4061	4949

在这种大背景下对我国目前的保险业进行研究，不论是对于促进我国保险业更加又好又快地发展，还是对于充分发挥保险业对于发展国民经济和改善居民生活的作用，都有着极为重

要的意义。

本章的研究目的如下：通过对我国的各个财产保险公司的基本情况进行各种分析，一方面找出构成财险公司基本特征的各变量之间的内在联系，另一方面找出各财险公司的共同特征或相异之处。

23.2 研究方法

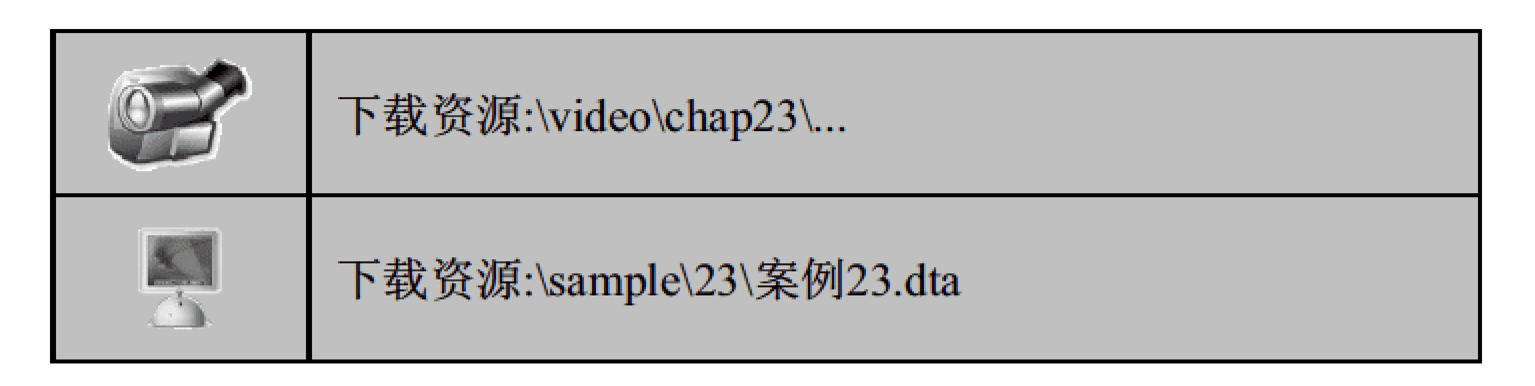
按照我国目前保险业的惯例，对于财产保险公司，可以用五个变量来描述其保险业务情况：保费收入、储金、赔案件数、赔款支出、未决赔款。其中，保费收入又按保险标的特点分为企业财产保险保费收入、机动车辆保险保费收入、货物运输保险保费收入、责任保险保费收入、信用保证保险保费收入、农业保险保费收入、短期健康保险保费收入、意外伤害保险保费收入、 其他保险保费收入 9 个组成部分；赔款支出按保险标的特点分为企业财产保险赔款支出、机动车辆保险赔款支出、货物运输保险赔款支出、责任保险赔款支出、信用保证保险赔款支出、农业保险赔款支出、短期健康保险赔款支出、意外伤害保险赔款支出、其他保险赔款支出 9 个组成部分。所以我们在进行分析研究的时候，考虑的关于保险业务的变量也与这些叙述相吻合。

本例采用的数据有《中国 2007 年各财产保险公司业务统计》《中国 2007 年各保险公司人员结构情况统计》等，这些数据都摘编自《中国保险年鉴 2008》。

采用的数据分析方法主要有描述性分析、相关分析、回归分析、因子分析、聚类分析等。

基本思路是：首先使用描述性分析来描述各个变量之间的基本特征，为后面的分析做好基础，然后使用相关分析、回归分析等研究保费收入、储金、赔案件数、赔款支出、未决赔款、公司总人数、人员构成等变量之间的关系；接着使用因子分析对构成保费收入和赔款支出的各个变量提取公因子；最后使用聚类分析依照人员构成特点和保费收入、赔款支出等变量对各财产保险公司进行聚类。

23.3 数据整理



	下载资源:\video\chap23\...
	下载资源:\sample\23\案例23.dta

因为本例采用的是现成的数据，所以根据第 1 章介绍的方法直接将所用数据录入 Stata 中即可。我们设置了 38 个变量，分别是“保险机构”“保费收入合计”“企业财产保险保费收入”“机动车辆保险保费收入”“货物运输保险保费收入”“责任保险保费收入”“信用保证保险保费收入”“农业保险保费收入”“短期健康保险保费收入”“意外伤害保险保费收入”“其他保险保费收入”“储金”“赔案件数”“赔款支出合计”“企业财产保险赔款支出”“机动车辆保险赔款支出”“货物运输保险赔款支出”“责任保险赔款支出”“信用保证保险赔款

支出”“农业保险赔款支出”“短期健康保险赔款支出”“意外伤害保险赔款支出”“其他保险赔款支出”“未决赔款”“总人数”“男”“女”“博士”“硕士”“学士”“大专”“中专以下”“高级”“中级”“初级”“三十五岁以下”“三十六岁到四十五岁”“四十六岁以上”等。其中，“保险机构”为字符串变量，其余变量均为数值型变量。我们把这 38 个变量分别定义为 V1-V38。样本是中国 2007 年各财产保险公司业务统计和人员构成的相关数据。录入完成后数据如图 23.1 所示。



	V1	V2	V3	V4	V5	V6	V7	V8
1	人保财险	88428.82	8867.3	62091.02	2978.79	3611.04	243.25	2657.7
2	国寿财险	789.19	71.81	639.66	8.93	28.7	.17	0
3	大地	10028.4	576.32	7669.14	171.47	243.11	22.93	10.5
4	中国保险	5483	809	3280	201	498	61	0
5	太平	3413.55	278.6	2686.61	115.99	79.6	-12.42	.08
6	中国信保	3240.77	0	0	0	0	3240.77	0
7	阳光财险	4153.46	279.14	3338.96	71.77	74.44	.78	0
8	中华联合	18342.07	670.32	14752.22	149.45	307.59	.09	1146.33
9	太保产险	23433.04	2796.61	16474.97	895.06	518.25	-87.06	6.91
10	平安产险	21449.53	2346.25	15165.45	641.81	615.76	88.56	6.73
11	华泰财险	2563.63	237.88	1245.99	196.24	138.67	2.02	0
12	天安	7371.4	350.03	6556.86	85.37	118.04	0	4.76
13	大众	1280.15	157.99	944.64	44.83	22.39	4.57	.56
14	华安	11301.9	50.95	619.97	10.49	11.89	-7.6	1.05
15	永安	5533.49	228.32	4766.02	36.67	113.34	16.31	.88
16	永诚	1505.61	438.94	838.29	11.76	37.74	.2	0
17	安信农险	276.62	22.35	33.41	.25	2.45	0	150.08
18	安邦	5752.21	116.16	5360.42	46.65	48.67	28.01	0
19	安华农险	1408.47	11.98	457.84	.7	2.65	0	889.11
20	天平汽车	0	0	1198.26	0	0	0	0
21	阳光农险	536.37	25.13	160.01	.09	4.48	0	320.98
22	渤海	742.46	34.54	642.38	4.57	9.89	0	0
23	都邦	2675.47	94.73	2331.39	34.55	31.17	0	0
24	华农	29.68	.72	26.43	.57	.37	0	0
25	民安	463.82	168.79	240.25	9.31	16.96	0	0
26	安诚	98.01	8.79	60.48	.42	2.15	.01	8.55
27	中银	524.08	111.72	261.45	18.94	21.44	0	0
28	中意财险	4.43	2.03	.41	.04	.04	0	0
29	美亚	833.21	139.6	2.28	187.27	342.32	8.61	0

图 23.1　数据 23

先做一下数据保存，然后展开后续分析。

23.4　描述性分析

本案例的数据变量除了城市这一字符串变量外都是定距变量，通过进行定距变量的基本描述性统计，我们可以得到数据的概要统计指标，包括平均值、最大值、最小值、标准差、百分位数、中位数、偏度系数和峰度系数等。我们通过获得这些指标，可以从整体上对拟分析的数据进行宏观把握，为后续进行更精深的数据分析做好必要准备。

23.4.1 Stata 分析过程

描述性分析的步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出“主界面”对话框。

02 在“主界面”对话框的“Command”文本框中输入命令：

summarize V2-V38,detail

03 设置完毕，按键盘上的回车键，等待输出结果。

23.4.2 结果分析

在 Stata 14.0“主界面”的结果窗口中，我们可以看到如图 23.2~图 23.20 所示的分析结果。

V2

	Percentiles	Smallest		
1%	0	0		
5%	4.66	4.43		
10%	29.68	4.66	Obs	42
25%	131.15	10.65	Sum of Wgt.	42
50%	639.415		Mean	5320.97
		Largest	Std. Dev.	14326.08
75%	4153.46	18342.07		
90%	11301.9	21449.53	Variance	2.05e+08
95%	21449.53	23433.04	Skewness	4.897387
99%	88428.82	88428.82	Kurtosis	28.42906

V3

	Percentiles	Smallest		
1%	0	0		
5%	.72	0		
10%	3.09	.72	Obs	42
25%	20.79	2.03	Sum of Wgt.	42
50%	70.425		Mean	462.7786
		Largest	Std. Dev.	1440.954
75%	237.88	809		
90%	670.32	2346.25	Variance	2076348
95%	2346.25	2796.61	Skewness	5.040446
99%	8867.3	8867.3	Kurtosis	29.16604

图 23.2 V2-V3 描述性分析结果图

V4

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	42
25%	.41	0	Sum of Wgt.	42
50%	250.85		Mean	3619.055
		Largest	Std. Dev.	10143.15
75%	2686.61	14752.22		
90%	7669.14	15165.45	Variance	1.03e+08
95%	15165.45	16474.97	Skewness	4.822536
99%	62091.02	62091.02	Kurtosis	27.7451

V5

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	.09	0	Obs	42
25%	2.84	.04	Sum of Wgt.	42
50%	19.91		Mean	155.0605
		Largest	Std. Dev.	477.5743
75%	128.52	201		
90%	196.24	641.81	Variance	228077.2
95%	641.81	895.06	Skewness	5.201632
99%	2978.79	2978.79	Kurtosis	30.71273

图 23.3 V4-V5 描述性分析结果图

V6

	Percentiles	Smallest		
1%	0	0		
5%	.04	0		
10%	.99	.04	Obs	42
25%	4.48	.37	Sum of Wgt.	42
50%	24.12		Mean	170.1857
		Largest	Std. Dev.	564.4741
75%	79.6	498		
90%	342.32	518.25	Variance	318631
95%	518.25	615.76	Skewness	5.599417
99%	3611.04	3611.04	Kurtosis	34.51707

V7

	Percentiles	Smallest		
1%	-87.06	-87.06		
5%	-7.6	-12.42		
10%	0	-7.6	Obs	42
25%	0	0	Sum of Wgt.	42
50%	.015		Mean	86.95524
		Largest	Std. Dev.	500.4107
75%	2.02	61		
90%	40.29	88.56	Variance	250410.9
95%	88.56	243.25	Skewness	6.175076
99%	3240.77	3240.77	Kurtosis	39.42169

图 23.4 V6-V7 描述性分析结果图

V8

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	42
25%	0	0	Sum of Wgt.	42
50%	0		Mean	123.946
		Largest	Std. Dev.	459.2832
75%	1.51	320.98		
90%	150.08	889.11	Variance	210941.1
95%	889.11	1146.33	Skewness	4.515681
99%	2657.7	2657.7	Kurtosis	23.95442

V9

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	42
25%	0	0	Sum of Wgt.	42
50%	0		Mean	54.58738
		Largest	Std. Dev.	187.1815
75%	7	277.48		
90%	29.36	370.58	Variance	35036.91
95%	370.58	422.84	Skewness	4.383199
99%	1079.69	1079.69	Kurtosis	23.10817

图 23.5 V8-V9 描述性分析结果图

V10

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	42
25%	2.2	0	Sum of Wgt.	42
50%	15.28		Mean	161.049
		Largest	Std. Dev.	362.7527
75%	124.53	579.07		
90%	517.24	712.12	Variance	131589.5
95%	712.12	1085.8	Skewness	3.540209
99%	1964.35	1964.35	Kurtosis	16.53628

V11

	Percentiles	Smallest		
1%	-11.55	-11.55		
5%	0	-10.83		
10%	0	0	Obs	42
25%	2.6	0	Sum of Wgt.	42
50%	20.985		Mean	266.0288
		Largest	Std. Dev.	821.4281
75%	109.51	680.22		
90%	540.21	1450.01	Variance	674744.1
95%	1450.01	1741.83	Skewness	4.734397
99%	4935.67	4935.67	Kurtosis	26.47074

图 23.6　V10-V11 描述性分析结果图

V12

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	42
25%	0	0	Sum of Wgt.	42
50%	0		Mean	623.7552
		Largest	Std. Dev.	2002.842
75%	0	1623.1		
90%	1069.01	6146.38	Variance	4330233
95%	6146.38	6518.06	Skewness	3.624817
99%	10492.66	10492.66	Kurtosis	15.43898

V13

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	42
25%	.08	0	Sum of Wgt.	42
50%	1.025		Mean	66.97833
		Largest	Std. Dev.	183.0587
75%	31.35	322.62		
90%	152.16	392.29	Variance	33510.48
95%	392.29	550.05	Skewness	3.649297
99%	975.93	975.93	Kurtosis	16.75716

图 23.7　V12-V13 描述性分析结果图

V14

	Percentiles	Smallest		
1%	-5.59	-5.59		
5%	0	0		
10%	.27	0	Obs	42
25%	22.34	.09	Sum of Wgt.	42
50%	124.17		Mean	2531.666
		Largest	Std. Dev.	7537.643
75%	1283.44	10823.26		
90%	4301	10861.02	Variance	5.68e+07
95%	10861.02	12693.56	Skewness	4.91368
99%	46230.73	46230.73	Kurtosis	28.38879

V15

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	.02	0	Obs	42
25%	3.57	.01	Sum of Wgt.	42
50%	20.785		Mean	240.9702
		Largest	Std. Dev.	849.8205
75%	82.59	453.52		
90%	275	1329.01	Variance	722194.8
95%	1329.01	1350.4	Skewness	5.277067
99%	5292.42	5292.42	Kurtosis	31.33531

图 23.8　V14-V15 描述性分析结果图

V16

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	42
25%	0	0	Sum of Wgt.	42
50%	32.765		Mean	1848.482
		Largest	Std. Dev.	5365.841
75%	969.19	7906.52		
90%	3809.56	8253.62	Variance	2.88e+07
95%	8253.62	10440.42	Skewness	4.648734
99%	32262.65	32262.65	Kurtosis	26.05442

V17

	Percentiles	Smallest		
1%	-.33	-.33		
5%	0	0		
10%	0	0	Obs	42
25%	.14	0	Sum of Wgt.	42
50%	6.18		Mean	59.05214
		Largest	Std. Dev.	193.6965
75%	40.76	111		
90%	101.58	190.79	Variance	37518.32
95%	190.79	333.87	Skewness	5.380819
99%	1220.28	1220.28	Kurtosis	32.38676

图 23.9　V16-V17 描述性分析结果图

V18

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	42
25%	.59	0	Sum of Wgt.	42
50%	3.39		Mean	69.381
		Largest	Std. Dev.	252.654
75%	14.62	194.87		
90%	157.59	260.03	Variance	63834.3
95%	260.03	388	Skewness	5.42957
99%	1589.81	1589.81	Kurtosis	32.84

V19

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	42
25%	0	0	Sum of Wgt.	42
50%	0		Mean	40.646
		Largest	Std. Dev.	165.158
75%	2.66	46.33		
90%	21.03	71.86	Variance	27277.
95%	71.86	680.69	Skewness	4.2901
99%	846.35	846.35	Kurtosis	19.7893

图 23.10　V18-V19 描述性分析结果图

V20

	Percentiles	Smallest		
1%	-43.82	-43.82		
5%	0	0		
10%	0	0	Obs	42
25%	0	0	Sum of Wgt.	42
50%	0		Mean	64.72619
		Largest	Std. Dev.	231.4957
75%	.65	123.01		
90%	4.81	695.53	Variance	53590.28
95%	695.53	864.91	Skewness	3.460775
99%	1060.29	1060.29	Kurtosis	13.49745

V21

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	42
25%	0	0	Sum of Wgt.	42
50%	0		Mean	43.25095
		Largest	Std. Dev.	162.2636
75%	4	136.4		
90%	21.39	204.64	Variance	26329.47
95%	284.64	335.87	Skewness	4.841413
99%	971.83	971.83	Kurtosis	27.18964

图 23.11　V20-V21 描述性分析结果图

V22

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	42
25%	.06	0	Sum of Wgt.	42
50%	1.59		Mean	58.27548
		Largest	Std. Dev.	131.3238
75%	42.78	176.83		
90%	163	300.48	Variance	17245.95
95%	300.48	458.73	Skewness	3.188049
99%	657.62	657.62	Kurtosis	13.34646

V23

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	42
25%	.44	0	Sum of Wgt.	42
50%	6.18		Mean	111.5288
		Largest	Std. Dev.	424.6677
75%	22.1	294.79		
90%	230.51	402	Variance	180342.7
95%	482	527.73	Skewness	5.558351
99%	2690.53	2690.53	Kurtosis	34.02718

图 23.12　V22-V23 描述性分析结果图

V24

	Percentiles	Smallest		
1%	0	0		
5%	.18	.05		
10%	.7	.18	Obs	42
25%	19.92	.29	Sum of Wgt.	42
50%	147.98		Mean	1342.067
		Largest	Std. Dev.	3655.938
75%	978.05	4803.18		
90%	2554	5650.9	Variance	1.34e+07
95%	5650.9	5868.6	Skewness	4.87945
99%	22519.96	22519.96	Kurtosis	28.27052

V25

	Percentiles	Smallest		
1%	30	30		
5%	49	35		
10%	62	49	Obs	42
25%	126	58	Sum of Wgt.	42
50%	1063		Mean	7497.167
		Largest	Std. Dev.	15377.93
75%	6955	29990		
90%	18806	33599	Variance	2.36e+08
95%	33599	60102	Skewness	2.870776
99%	70849	70849	Kurtosis	10.98637

图 23.13　V24-V25 描述性分析结果图

V26

	Percentiles	Smallest		
1%	15	15		
5%	18	17		
10%	26	18	Obs	42
25%	72	19	Sum of Wgt.	42
50%	616.5		Mean	4088.81
		Largest	Std. Dev.	8867.813
75%	3743	15062		
90%	9008	18170	Variance	7.86e+07
95%	18170	37549	Skewness	3.09603
99%	40217	40217	Kurtosis	12.19519

V27

	Percentiles	Smallest		
1%	13	13		
5%	30	17		
10%	38	30	Obs	42
25%	64	31	Sum of Wgt.	42
50%	466		Mean	3406.381
		Largest	Std. Dev.	6691.093
75%	3212	14928		
90%	9894	15429	Variance	4.48e+07
95%	15429	19885	Skewness	2.816371
99%	33300	33300	Kurtosis	11.43047

图 23.14　V26-V27 描述性分析结果图

V28

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	42
25%	0	0	Sum of Wgt.	42
50%	2.5		Mean	5.952381
		Largest	Std. Dev.	10.86267
75%	5	15		
90%	14	32	Variance	117.9977
95%	32	39	Skewness	2.882395
99%	52	52	Kurtosis	11.09974

V29

	Percentiles	Smallest		
1%	2	2		
5%	7	2		
10%	7	7	Obs	42
25%	17	7	Sum of Wgt.	42
50%	48.5		Mean	290.4286
		Largest	Std. Dev.	1017.977
75%	163	456		
90%	409	757	Variance	1036278
95%	757	1204	Skewness	5.80614
99%	6571	6571	Kurtosis	36.22177

图 23.15　V28-V29 描述性分析结果图

V30

	Percentiles	Smallest		
1%	21	21		
5%	28	26		
10%	36	28	Obs	42
25%	71	34	Sum of Wgt.	42
50%	409		Mean	2513.5
		Largest	Std. Dev.	5850.047
75%	2476	7475		
90%	6356	7821	Variance	3.42e+07
95%	7821	21979	Skewness	3.683347
99%	30738	30738	Kurtosis	16.65388

V31

	Percentiles	Smallest		
1%	0	0		
5%	5	4		
10%	8	5	Obs	42
25%	23	8	Sum of Wgt.	42
50%	349		Mean	3002.976
		Largest	Std. Dev.	5895.036
75%	2797	12348		
90%	8048	13195	Variance	3.48e+07
95%	13195	23989	Skewness	2.653177
99%	25426	25426	Kurtosis	9.683805

图 23.16　V30-V31 描述性分析结果图

V32

	Percentiles	Smallest		
1%	0	0		
5%	1	0		
10%	2	1	Obs	42
25%	6	1	Sum of Wgt.	42
50%	81		Mean	1682.333
		Largest	Std. Dev.	3438.66
75%	1436	8062		
90%	6648	10140	Variance	1.18e+07
95%	10140	12891	Skewness	2.348265
99%	13305	13305	Kurtosis	7.445075

V33

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	42
25%	0	0	Sum of Wgt.	42
50%	15.5		Mean	104.7619
		Largest	Std. Dev.	288.4233
75%	86	231		
90%	192	252	Variance	83187.99
95%	252	910	Skewness	4.447432
99%	1668	1668	Kurtosis	23.09082

图 23.17　V32-V33 描述性分析结果图

V34

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	42
25%	1	0	Sum of Wgt.	42
50%	85.5		Mean	847.9524
		Largest	Std. Dev.	2617.983
75%	562	2136		
90%	1446	2972	Variance	6853833
95%	2972	4908	Skewness	5.159856
99%	16296	16296	Kurtosis	30.45409

V35

	Percentiles	Smallest		
1%	0	0		
5%	0	0		
10%	0	0	Obs	42
25%	0	0	Sum of Wgt.	42
50%	57		Mean	1303.881
		Largest	Std. Dev.	3985.832
75%	570	2791		
90%	1276	7643	Variance	1.59e+07
95%	7643	15733	Skewness	3.81126
99%	20039	20039	Kurtosis	16.6555

图 23.18　V34-V35 描述性分析结果图

V36

	Percentiles	Smallest		
1%	23	23		
5%	34	25		
10%	49	34	Obs	42
25%	92	47	Sum of Wgt.	42
50%	716		Mean	4504.643
		Largest	Std. Dev.	9600.217
75%	4024	17295		
90%	13206	19189	Variance	9.22e+07
95%	19189	19757	Skewness	3.707457
99%	54491	54491	Kurtosis	18.79216

V37

	Percentiles	Smallest		
1%	4	4		
5%	10	8		
10%	11	10	Obs	42
25%	16	11	Sum of Wgt.	42
50%	254		Mean	2181.976
		Largest	Std. Dev.	4819.379
75%	1907	8742		
90%	5544	10576	Variance	2.32e+07
95%	10576	14209	Skewness	3.357023
99%	25559	25559	Kurtosis	15.15075

图 23.19　V36-V37 描述性分析结果图

V38

	Percentiles	Smallest		
1%	0	0		
5%	1	0		
10%	1	1	Obs	42
25%	5	1	Sum of Wgt.	42
50%	53		Mean	808.4524
		Largest	Std. Dev.	2702.977
75%	656	1491		
90%	1491	2149	Variance	7306087
95%	2149	3834	Skewness	5.575903
99%	17248	17248	Kurtosis	34.23351

图 23.20　V38 描述性分析结果图

在图 23.2~图 23.20 所示的分析结果中，我们可以得到很多信息。此处限于篇幅不再针对各个变量一一展开说明，以变量 V38 为例进行解释。信息包括：

（1）百分位数（Percentiles）

可以看出变量 V38 的第一个四分位数（25%）是 5，第二个四分位数（50%）是 53。

（2）四个最小值（Smallest）

变量 V38 最小的四个数据值分别是 0、0、1、1。

（3）四个最大值（Largest）

变量 V38 最大的四个数据值分别是 1491、2149、3834、17248。

（4）平均值（Mean）和标准差（Variance）

变量 V38 的平均值为 808.4524，标准差是 2702.977。

（5）偏度（Skewness）和峰度（Kurtosis）

变量 V38 的偏度为 5.575903，为正偏度。

变量 V38 的峰度为 34.23351，有一个比正态分布更长的尾巴。

从上面的描述性分析结果中，我们可以比较轻松地看出，所有数据中没有极端数据，数据间的量纲差距也在可接受范围之内，可以进入下一步的分析过程。

23.5 相关分析

对于相关分析，我们准备进行以下几个部分。

第一，对“保费收入合计”的 9 个组成部分——“企业财产保险保费收入”“机动车辆保险保费收入”“货物运输保险保费收入”“责任保险保费收入”“信用保证保险保费收入”“农业保险保费收入”“短期健康保险保费收入”“意外伤害保险保费收入”“其他保险保费收入”进行简单相关分析。

第二，对“赔款支出合计”的 9 个组成部分——“企业财产保险赔款支出”“机动车辆保险赔款支出”“货物运输保险赔款支出”“责任保险赔款支出”“信用保证保险赔款支出”“农业保险赔款支出”“短期健康保险赔款支出”“意外伤害保险赔款支出”“其他保险赔款支出”进行简单相关分析。

第三，对“保费收入合计”“赔款支出合计”“总人数”这 3 个变量进行简单相关分析。

第四，对“赔案件数”“赔款支出合计”“未决赔款”这 3 个变量进行简单相关分析。

1. 对“保费收入合计”的 9 个组成部分进行简单相关分析

操作步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出“主界面”对话框。

02 在“主界面”对话框的“Command”文本框中输入命令：

（1）correlate V3-V11

本命令旨在使用简单相关分析方法研究 V3~V11 这 9 个变量之间的相关关系。

（2）pwcorr V3-V11,sidak sig star(0.01)

本命令旨在判断 V3~V11 这 9 个变量之间的相关性在置信水平为 99%时是否显著。

03 设置完毕，按键盘上的回车键，等待输出结果。

结果分析如图 23.21、图 23.22 所示。从图 23.21 可以看出，构成“保费收入合计”的 9 个组成部分除“信用保证保险保费收入”(V7)与别的变量相关关系较弱外，其他变量之间都具有很强的相关性。

```
. correlate V3-V11
(obs=42)
```

	V3	V4	V5	V6	V7	V8	V9	V10	V11
V3	1.0000								
V4	0.9775	1.0000							
V5	0.9927	0.9667	1.0000						
V6	0.9774	0.9655	0.9803	1.0000					
V7	0.0170	0.0114	0.0160	0.0267	1.0000				
V8	0.8105	0.8580	0.8111	0.8541	0.0206	1.0000			
V9	0.8983	0.9366	0.8854	0.9157	0.0280	0.8558	1.0000		
V10	0.9433	0.9593	0.9245	0.8946	-0.0148	0.7468	0.8704	1.0000	
V11	0.9914	0.9667	0.9868	0.9632	0.0148	0.7890	0.8924	0.9447	1.0000

图 23.21 相关分析结果图 1

```
. pwcorr V3-V11,sidak sig star(0.01)
```

	V3	V4	V5	V6	V7	V8	V9
V3	1.0000						
V4	0.9775*	1.0000					
	0.0000						
V5	0.9927*	0.9667*	1.0000				
	0.0000	0.0000					
V6	0.9774*	0.9655*	0.9803*	1.0000			
	0.0000	0.0000	0.0000				
V7	0.0170	0.0114	0.0160	0.0267	1.0000		
	1.0000	1.0000	1.0000	1.0000			
V8	0.8105*	0.8580*	0.8111*	0.8541*	0.0206	1.0000	
	0.0000	0.0000	0.0000	0.0000	1.0000		
V9	0.8983*	0.9366*	0.8854*	0.9157*	0.0280	0.8558*	1.0000
	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	
V10	0.9433*	0.9593*	0.9245*	0.8946*	-0.0148	0.7468*	0.8704*
	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000
V11	0.9914*	0.9667*	0.9868*	0.9632*	0.0148	0.7890*	0.8924*
	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000

	V10	V11
V10	1.0000	
V11	0.9447*	1.0000
	0.0000	

图 23.22 相关分析结果图 2

从图 23.22 中可以看出，绝大多数变量之间相关性在 1%的显著性水平上显著。

2. 对“赔款支出合计”的 9 个组成部分进行简单相关分析

操作步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出“主界面”对话框。

02 在“主界面”对话框的“Command”文本框中输入命令：

（1）correlate V15-V23

本命令旨在使用简单相关分析方法研究 V15~V23 这 9 个变量之间的相关关系。

（2）pwcorr V15-V23,sidak sig star(0.01)

本命令旨在判断 V15-V23 这 9 个变量之间的相关性在置信水平为 99%时是否显著。

03 设置完毕，按键盘上的回车键，等待输出结果。

结果分析如图 23.23~图 23.24 所示。从图 23.23 可以看出，构成“赔款支出合计”的 9 个组成部分变量之间都具有比较强的相关性。

. correlate V15-V23
(obs=42)

	V15	V16	V17	V18	V19	V20	V21	V22	V23
V15	1.0000								
V16	0.9685	1.0000							
V17	0.9881	0.9457	1.0000						
V18	0.9697	0.9412	0.9708	1.0000					
V19	0.5835	0.5476	0.5802	0.5820	1.0000				
V20	0.5280	0.5975	0.5122	0.5418	0.3070	1.0000			
V21	0.9401	0.9596	0.9067	0.9301	0.5584	0.5994	1.0000		
V22	0.8744	0.9253	0.8575	0.8161	0.4366	0.5230	0.8058	1.0000	
V23	0.9771	0.9495	0.9842	0.9920	0.5872	0.5561	0.9177	0.8462	1.0000

图 23.23　相关分析结果图 3

. pwcorr V15-V23,sidak sig star(0.01)

	V15	V16	V17	V18	V19	V20	V21
V15	1.0000						
V16	0.9685*	1.0000					
	0.0000						
V17	0.9881*	0.9457*	1.0000				
	0.0000	0.0000					
V18	0.9697*	0.9412*	0.9708*	1.0000			
	0.0000	0.0000	0.0000				
V19	0.5835*	0.5476*	0.5802*	0.5820*	1.0000		
	0.0018	0.0063	0.0020	0.0019			
V20	0.5280	0.5975*	0.5122	0.5418*	0.3070	1.0000	
	0.0117	0.0011	0.0187	0.0076	0.8298		
V21	0.9401*	0.9596*	0.9067*	0.9301*	0.5584*	0.5994*	1.0000
	0.0000	0.0000	0.0000	0.0000	0.0044	0.0010	
V22	0.8744*	0.9253*	0.8575*	0.8161*	0.4366	0.5230	0.8058*
	0.0000	0.0000	0.0000	0.0000	0.1295	0.0136	0.0000
V23	0.9771*	0.9495*	0.9842*	0.9920*	0.5872*	0.5561*	0.9177*
	0.0000	0.0000	0.0000	0.0000	0.0016	0.0047	0.0000

	V22	V23
V22	1.0000	
V23	0.8462*	1.0000
	0.0000	

图 23.24　相关分析结果图 4

从图 23.24 中可以看出，大部分变量的相关性很强，在 0.01 的显著性水平上显著。

3. 对“保费收入合计”“赔款支出合计”“总人数”这 3 个变量进行简单相关分析

操作步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出“主界面”对话框。

02 在“主界面”对话框的“Command”文本框中输入命令：

（1）correlate V2 V14 V25

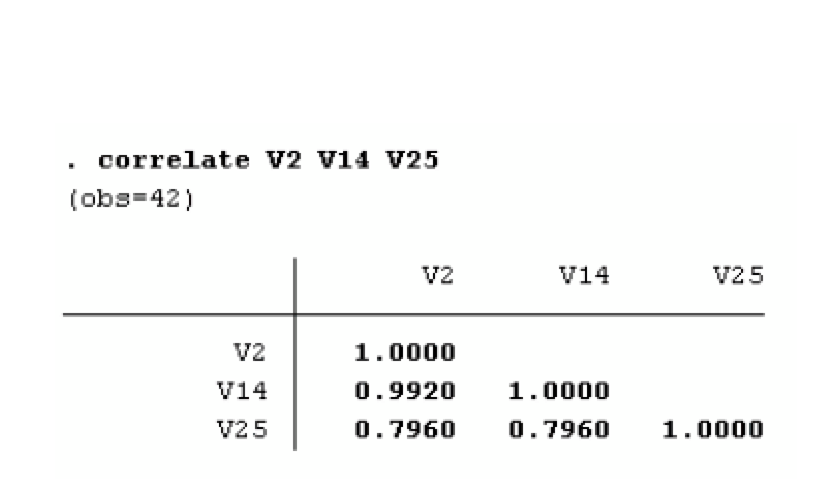
本命令旨在使用简单相关分析方法研究 V2、V14、V25 这 3 个变量之间的相关关系。

（2）pwcorr V2 V14 V25,sidak sig star(0.01)

本命令旨在判断 V2、V14、V25 这 3 个变量之间的相关性在置信水平为 99%时是否显著。

03 设置完毕，按键盘上的回车键，等待输出结果。

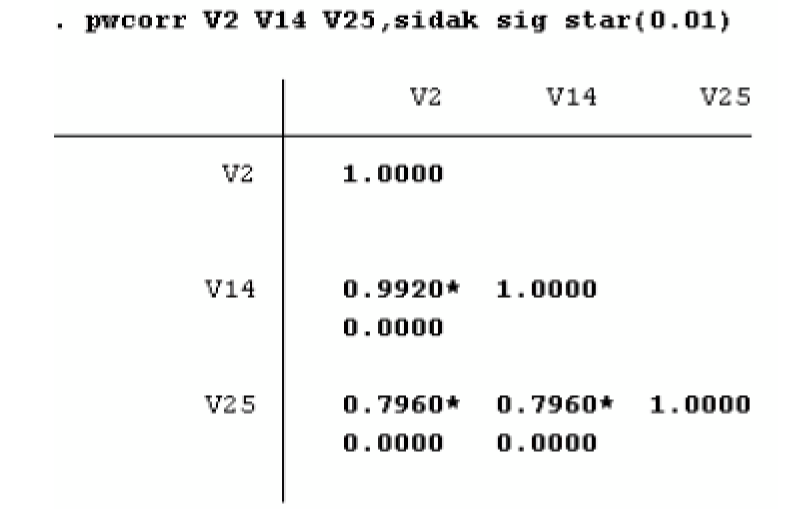
结果分析如图 23.25、图 23.26 所示。从图 23.25 可以看出，“保费收入合计”“赔款支出合计”“总人数”这三个变量之间具有很强的相关性。



. correlate V2 V14 V25
(obs=42)

	V2	V14	V25
V2	1.0000		
V14	0.9920	1.0000	
V25	0.7960	0.7960	1.0000

图 23.25 相关分析结果图 5



. pwcorr V2 V14 V25,sidak sig star(0.01)

	V2	V14	V25
V2	1.0000		
V14	0.9920*	1.0000	
	0.0000		
V25	0.7960*	0.7960*	1.0000
	0.0000	0.0000	

图 23.26 相关分析结果图 6

从图 23.26 中可以看出，“保费收入合计”“赔款支出合计”“总人数”这 3 个变量之间的相关性在 1%的显著性水平上显著。

4. 对“赔案件数”“赔款支出合计”“未决赔款”这 3 个变量进行简单相关分析

操作步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出“主界面”对话框。

02 在“主界面”对话框的“Command”文本框中输入命令：

（1）correlate V13 V14 V24

本命令旨在使用简单相关分析方法研究 V13、V14、V24 这 3 个变量之间的相关关系。

（2）pwcorr V13 V14 V24,sidak sig star(0.01)

本命令旨在判断 V13、V14、V24 这 3 个变量之间的相关性在置信水平为 99%时是否显著。

03 设置完毕，按键盘上的回车键，等待输出结果。

结果分析如图 23.27、图 23.28 所示。从图 23.27 可以看出，“赔案件数”“赔款支出合计”“未决赔款”这 3 个变量之间具有很强的相关性。

. correlate V13 V14 V24
(obs=42)

	V13	V14	V24
V13	1.0000		
V14	0.9513	1.0000	
V24	0.9435	0.9930	1.0000

图 23.27 相关分析结果图 7

. pwcorr V13 V14 V24,sidak sig star(0.01)

	V13	V14	V24
V13	1.0000		
V14	0.9513*	1.0000	
	0.0000		
V24	0.9435*	0.9930*	1.0000
	0.0000	0.0000	

图 23.28 相关分析结果图 8

从图 23.28 中可以看出，“赔案件数”“赔款支出合计”“未决赔款”这 3 个变量之间的相关性在 1%的显著性水平上显著。

23.6 回归分析

对于回归分析，我们准备进行以下几个部分。

第一，以“保费收入合计”为因变量，以“男”“女”“博士”“硕士”“学士”“大专”“中专以下”“高级”“中级”“初级”“三十五岁以下”“三十六岁到四十五岁”“四十六岁以上”为自变量，进行最小二乘线性回归。

第二，以“赔款支出合计”为因变量，以“男”“女”“博士”“硕士”“学士”“大专”“中专以下”“高级”“中级”“初级”“三十五岁以下”“三十六岁到四十五岁”“四十六岁以上”为自变量，进行最小二乘线性回归。

1. 以“保费收入合计”为因变量，以“男”“女”“博士”“硕士”“学士”“大专”“中专以下”“高级”“中级”“初级”“三十五岁以下”“三十六岁到四十五岁”“四十六岁以上”为自变量，进行最小二乘回归

建立线性模型：

V2=a*V26+b*V27+c*V28+d*V29+e*V30+f*V31+g*V32+h*V33+i*V34+j*V35+k*V36+l*V37+m*V38+ u

普通最小二乘回归分析步骤及结果如下：

01 进入 Stata 14.0，打开相关数据文件，弹出“主界面”对话框。

02 在“主界面”对话框的“Command”文本框中输入命令：

（1）sw regress V2 V26-V38,pr(0.05)

本命令的含义是使用逐步回归分析方法，以“保费收入合计”为因变量，以“男”“女”“博士”“硕士”“学士”“大专”“中专以下”“高级”“中级”“初级”“三十五岁以下”“三十六岁到四十五岁”“四十六岁以上”为自变量，进行最小二乘回归分析。

（2）encode V2,gen(company)

本命令旨在将 V2 这一字符串变量转化为数值型变量 company，以便进行下一步操作。

（3）reg V2 V26-V38,vce(cluster company)

本命令的含义是以“保费收入合计”为因变量，以“男”“女”“博士”“硕士”“学士”“大专”“中专以下”“高级”“中级”“初级”“三十五岁以下”“三十六岁到四十五岁”“四十六岁以上”为自变量，并使用以“bank”为聚类变量的聚类稳健标准差，进行最小二乘回归分析。

（4）reg V2 V26-V30 V33-V38,vce(cluster company) nocon

本命令是在上步回归的基础上，剔除掉不显著的自变量以后，以“保费收入合计”为因变量，以 V26-V30、V33-V38 等变量为自变量，并使用以“company”为聚类变量的聚类稳健标准差，进行最小二乘回归分析。

03 设置完毕，按键盘上的回车键确认。

在 Stata1 4.0“主界面”的结果窗口可以看到如图 23.29~图 23.32 所示的分析结果：

（1）图 23.29 是使用逐步回归分析方法，以“保费收入合计”为因变量，以“男”“女”“博士”“硕士”“学士”“大专”“中专以下”“高级”“中级”“初级”“三十五岁以下”“三十六岁到四十五岁”“四十六岁以上”为自变量，进行最小二乘回归分析的结果。

```
. sw regress V2  V26-V38,pr(0.05)
                      begin with full model
p = 0.9997 >= 0.0500  removing V30
p = 0.8471 >= 0.0500  removing V37
p = 0.1990 >= 0.0500  removing V35
p = 0.2081 >= 0.0500  removing V29

      Source |       SS       df       MS              Number of obs =      42
-------------+------------------------------           F(  9,     32) =  668.20
       Model |  8.3702e+09     9   930018468           Prob > F      =  0.0000
    Residual |  44538456.9    32  1391826.78           R-squared     =  0.9947
-------------+------------------------------           Adj R-squared =  0.9932
       Total |  8.4147e+09    41   205236699           Root MSE      =  1179.8

------------------------------------------------------------------------------
          V2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         V26 |  -4.676808   1.199748    -3.90   0.000    -7.120615     -2.233
         V27 |  -9.540191   1.456879    -6.55   0.000    -12.50776  -6.572626
         V28 |   316.9149    75.5445     4.20   0.000     163.0358    470.794
         V36 |   6.465953    1.16867     5.53   0.000      4.08545   8.846457
         V38 |   7.575736   1.477396     5.13   0.000     4.566379   10.58509
         V31 |    4.26689   .6656031     6.41   0.000     2.911101   5.622679
         V32 |   2.087525   .4284813     4.87   0.000     1.214737   2.960312
         V33 |  -49.23041   10.12065    -4.86   0.000    -69.84549  -28.61533
         V34 |    10.0708   2.207833     4.56   0.000     5.573595   14.56801
       _cons |   5.321901   262.0074     0.02   0.984    -528.3698   539.0136
------------------------------------------------------------------------------
```

图 23.29 回归分析结果图 1

在上述分析结果中，我们可以得到很多信息。可以看出共有 42 个样本参与了分析，模型的 F 值(9, 32) =668.20，P 值（Prob > F） = 0.0000，说明模型整体上是非常显著的。模型的可决系数（R-squared）为 0.9947，模型修正的可决系数（Adj R-squared）为 0.9932，说明模型的解释能力是非常优秀接近完美的。

模型经过四次剔除变量后得到最终结果。第一个模型是包含全部自变量的全模型，该模型中 V30 变量的系数显著性 P 值高达 0.9997，被剔除掉；第二个模型是剔除掉自变量 V30 以后的模型，该模型中 V37 变量的系数显著性 P 值高达 0.8471，被剔除掉；第三个模型是剔除掉自变量 V30、V37 以后的模型，该模型中 V35 变量的系数显著性 P 值高达 0.1990，被剔除掉；第四个模型是剔除掉自变量 V30、V37、V35 以后的模型，该模型中 V29 变量的系数显著性 P 值高达 0.2081，被剔除掉。剔除掉自变量 V30、V37、V35、V29 以后，我们得到最终回归模型。

在最终回归模型中，变量 V26 的系数标准误是 1.199748，t 值为-3.90，P 值为 0.000，系数是非常显著的，95%的置信区间为[-7.120615, -2.233]。变量 V27 的系数标准误是 1.456879，t 值为-6.55，P 值为 0.000，系数是非常显著的，95%的置信区间为[-12.50776, -6.572626]。变量 V28 的系数标准误是 75.5445，t 值为 4.20，P 值为 0.000，系数是非常显著的，95%的置信区间为[163.0358, 470.794]。变量 V36 的系数标准误是 1.16867，t 值为 5.53，P 值为 0.000，系数是非常显著的，95%的置信区间为[4.08545, 8.846457]。变量 V38 的系数标准误是 1.477396，t 值为 5.13，P 值为 0.000，系数是非常显著的，95%的置信区间为[4.566379, 10.58509]。变量 V31 的系数标准误是 0.6656031，t 值为 6.41，P 值为 0.000，系数是非常显著的，95%的置信区间为[2.911101, 5.622679]。变量 V32 的系数标准误是 0.4284813，t 值为 4.87，P 值为 0.000，

系数是非常显著的，95%的置信区间为[1.214737, 2.960312]。变量 V34 的系数标准误是 2.207833，t 值为 4.56，P 值为 0.000，系数是非常显著的，95%的置信区间为[5.573595, 14.56801]。常数项的系数标准误是 262.0074，t 值为 0.02，P 值为 0.984，系数是非常不显著的，95%的置信区间为[-528.3698, 539.0136]。

最终最小二乘回归模型的方程是：

保费收入合计= -4.676808 *男-9.540191 *女+ 316.9149 *V 博士+ 4.26689 *大专+ 2.087525 *中专以下-49.23041 *高级+ 10.0708 *中级+ 6.465953 *三十五岁以下+ 7.575736 *四十六岁以上+ u

经过以上最小二乘回归分析，可以发现我国财产保险公司的总保费收入水平与公司职员的性别、年龄、职称、文化水平都有一定的显著关系。具体而言，中级职称或者大专、中专以下、博士学历或者三十五岁以下、四十六岁以上的职员对公司的总保费收入有拉动效应，尤其是博士学历的职员，每增加一单位会带来对应保费收入的 300 多倍的增加；高级职称或者男性、女性的职员对公司的总保费收入有拖后效应。

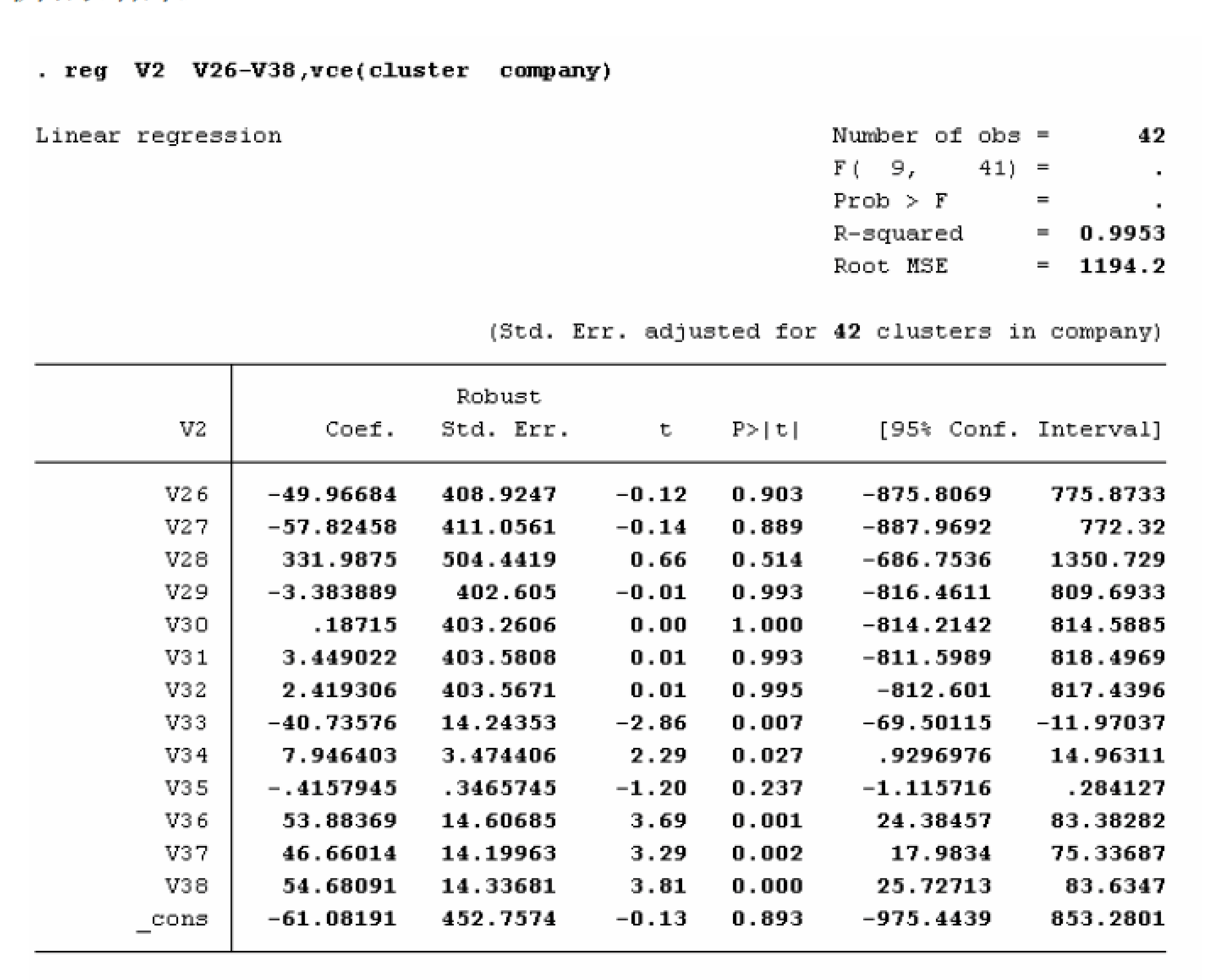
（2）图 23.30 是将 V2 这一字符串变量转化为数值型变量 company 的结果。选择“Data” | “Data Editor” | “Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 23.30 所示的变量 company 的相关数据。

	V32	V33	V34	V35	V36	V37	V38	company
1	12891	1668	16296	20039	17295	25559	17248	人保财险
2	544	76	562	570	2743	1622	324	国寿财险
3	2866	115	1419	925	9022	5022	1491	大地
4	3769	231	1446	1276	13206	4680	920	中国保险
5	1740	76	524	654	4024	1907	1024	太平
6	75	104	225	128	918	231	64	中国信保
7	1436	86	683	473	3752	1960	443	阳光财险
8	10140	192	2136	2791	19757	8742	1491	中华联合
9	13305	252	2972	15733	19189	10576	3834	太保产险
10	8062	910	4908	7643	54491	14209	2149	平安产险
11	326	23	260	199	1539	663	146	华泰财险
12	6648	116	1065	1013	12034	5544	1157	天安
13	639	8	131	108	914	530	249	大众
14	1073	17	165	152	6932	1151	255	华安
15	4047	86	896	1253	7852	3518	1010	永安
16	14	14	138	56	436	228	28	永诚
17	13	9	37	27	51	50	25	安信农险
18	1483	186	870	1127	6586	2227	685	安邦
19	453	60	156	95	896	370	338	安华农险
20	67	26	81	134	669	116	31	天平汽车
21	8	40	59	37	89	117	45	阳光农险
22	311	19	195	58	1133	787	35	渤海
23	103	0	0	0	1992	518	81	都邦
24	10	8	35	7	95	44	11	华农
25	219	29	205	86	774	277	51	民安
26	150	27	90	51	120	321	656	安诚
27	87	0	0	0	666	308	55	中银
28	1	2	3	1	49	14	2	中意财险
29	74	6	0	0	763	155	61	美亚

图 23.30　回归分析结果图 2

（3）图 23.31 是以“保费收入合计”为因变量，以“男”“女”“博士”“硕士”“学士”“大专”“中专以下”“高级”“中级”“初级”“三十五岁以下”“三十六岁到四十五岁”“四十六岁以上”为自变量，并使用以“company”为聚类变量的聚类稳健标准差，进行

最小二乘回归分析的结果。



```
. reg  V2  V26-V38,vce(cluster  company)

Linear regression                               Number of obs =       42
                                                F(  9,     41) =        .
                                                Prob > F       =        .
                                                R-squared      =   0.9953
                                                Root MSE       =   1194.2

                             (Std. Err. adjusted for 42 clusters in company)
```

V2	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
V26	-49.96684	408.9247	-0.12	0.903	-875.8069	775.8733
V27	-57.82458	411.0561	-0.14	0.889	-887.9692	772.32
V28	331.9875	504.4419	0.66	0.514	-686.7536	1350.729
V29	-3.383889	402.605	-0.01	0.993	-816.4611	809.6933
V30	.18715	403.2606	0.00	1.000	-814.2142	814.5885
V31	3.449022	403.5808	0.01	0.993	-811.5989	818.4969
V32	2.419306	403.5671	0.01	0.995	-812.601	817.4396
V33	-40.73576	14.24353	-2.86	0.007	-69.50115	-11.97037
V34	7.946403	3.474406	2.29	0.027	.9296976	14.96311
V35	-.4157945	.3465745	-1.20	0.237	-1.115716	.284127
V36	53.88369	14.60685	3.69	0.001	24.38457	83.38282
V37	46.66014	14.19963	3.29	0.002	17.9834	75.33687
V38	54.68091	14.33681	3.81	0.000	25.72713	83.6347
_cons	-61.08191	452.7574	-0.13	0.893	-975.4439	853.2801

图 23.31　回归分析结果图 3

我们可以看出，该结果中有很多变量系数的显著性是非常差的，需要把不显著的变量进行剔除后再进行进一步分析。

（4）图 23.32 是在上步回归的基础上，剔除掉不显著的自变量以后，以“保费收入合计”为因变量，以 V26~V30、V33~V38 为自变量，并使用以“company”为聚类变量的聚类稳健标准差，进行最小二乘回归分析的结果。

```
. reg  V2  V26-V30 V33-V38,vce(cluster  company) nocon

Linear regression                               Number of obs =       42
                                                F(  8,     41) =        .
                                                Prob > F       =        .
                                                R-squared      =   0.9956
                                                Root MSE       =   1161.6

                             (Std. Err. adjusted for 42 clusters in company|
```

V2	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
V26	-59.62343	20.40121	-2.92	0.006	-100.8245	-18.42236
V27	-68.47514	22.08822	-3.10	0.003	-113.0832	-23.86708
V28	317.2921	121.9827	2.60	0.013	70.94298	563.6413
V29	-7.326464	2.053083	-3.57	0.001	-11.47275	-3.180179
V30	-2.483157	1.085676	-2.29	0.027	-4.675724	-.2905903
V33	-37.78445	12.01396	-3.15	0.003	-62.04714	-13.52175
V34	7.574484	3.055632	2.48	0.017	1.403511	13.74546
V35	-.667158	.236078	-2.83	0.007	-1.143927	-.1903888
V36	67.18056	21.65574	3.10	0.003	23.4459	110.9152
V37	59.41212	20.86529	2.85	0.007	17.27382	101.5504
V38	67.9182	21.51962	3.16	0.003	24.45843	111.378

图 23.32　回归分析结果图 4

可以看出，在剔除掉不显著的自变量以后，以“保费收入合计”为因变量，以 V26~V30、V33~V38 为自变量，并使用以“company”为聚类变量的聚类稳健标准差，进行最小二乘回归分析的结果与普通最小二乘回归分析有所区别。

在该模型中，最终保留的自变量有“男”“女”“博士”“硕士”“学士”“高级”“中

级”“初级”“三十五岁以下”“三十六岁到四十五岁”“四十六岁以上”。

该模型方程为：

保费收入合计= -59.62343 *男-68.47514 *女+ 317.2921 *V 博士-7.326464 *硕士-2.483157 *学士-37.78445 *高级+ 7.574484 *中级-0.667158 *初级+ 67.18056 *三十五岁以下+ 59.41212 *三十六岁到四十五岁+ 67.9182 *四十六岁以上+ u

经过以上分析，可以发现我国财产保险公司的总保费收入水平与公司职员的性别、年龄、职称、文化水平都有一定的显著关系。具体而言，中级职称或者大专、中专以下、博士学历或者三十五岁以下、三十六岁到四十五岁、四十六岁以上的职员对公司的总保费收入有拉动效应，尤其是博士学历的职员，每增加一单位会带来对应保费收入的 300 多倍的增加；高级职称、初级职称或者硕士学历、学士学历或者男性、女性的职员对公司的总保费收入有拖后效应。

2. 以“赔款支出合计”为因变量，以“男”“女”“博士”“硕士”“学士”“大专”“中专以下”“高级”“中级”“初级”“三十五岁以下”“三十六岁到四十五岁”“四十六岁以上”为自变量，进行最小二乘回归

建立线性模型：

V14=a*V26+b*V27+c*V28+d*V29+e*V30+f*V31+g*V32+h*V33+i*V34+j*V35+k*V36+l*V37+m*V38+ u

普通最小二乘回归分析步骤及结果如下：

01 进入 Stata 14.0，打开相关数据文件，弹出“主界面”对话框。

02 在“主界面”对话框的“Command”文本框中输入命令：

（1）sw regress V14　V26-V38,pr(0.05)

本命令的含义是使用逐步回归分析方法，以“赔款支出合计”为因变量，以“男”“女”“博士”“硕士”“学士”“大专”“中专以下”“高级”“中级”“初级”“三十五岁以下”“三十六岁到四十五岁”“四十六岁以上”为自变量，进行最小二乘回归分析。

（2）reg　V14　V26-V38,vce(cluster　company)

本命令的含义是以“赔款支出合计”为因变量，以“男”“女”“博士”“硕士”“学士”“大专”“中专以下”“高级”“中级”“初级”“三十五岁以下”“三十六岁到四十五岁”“四十六岁以上”为自变量，并使用以“bank”为聚类变量的聚类稳健标准差，进行最小二乘回归分析。

（3）reg　V14　V27　V29-V31　V34-V38,vce(cluster　company) nocon

本命令是在上步回归的基础上，剔除掉不显著的自变量以后，以“赔款支出合计”为因变量，以 V26~V30、V33~V38 为自变量，并使用以“company”为聚类变量的聚类稳健标准差，进行最小二乘回归分析。

03 设置完毕，按键盘上的回车键确认。

在 Stata 14.0“主界面”的结果窗口我们可以看到如图 23.33~图 23.35 所示的分析结果：

（1）图 23.33 是使用逐步回归分析方法，以“赔款支出合计”为因变量，以“男”“女”

“博士”“硕士“学士”“大专”“中专以下”“高级”“中级”“初级”“三十五岁以下”“三十六岁到四十五岁”“四十六岁以上”为自变量，进行最小二乘回归分析的结果。

```
. sw regress V14  V26-V38,pr(0.05)
                      begin with full model
p = 0.9855 >= 0.0500  removing V26
p = 0.9251 >= 0.0500  removing V32
p = 0.4340 >= 0.0500  removing V33
p = 0.1746 >= 0.0500  removing V28

      Source |       SS       df       MS              Number of obs =      42
-------------+------------------------------           F(  9,    32) =  708.01
       Model |  2.3178e+09     9   257535399           Prob > F      =  0.0000
    Residual |  11639820.5    32  363744.391           R-squared     =  0.9950
-------------+------------------------------           Adj R-squared =  0.9936
       Total |  2.3295e+09    41  56816058.8           Root MSE      =  603.11

------------------------------------------------------------------------------
         V14 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         V38 |   2.330955   .8543404     2.73   0.010     .5907202    4.071189
         V27 |  -3.040764   .8334345    -3.65   0.001    -4.738415   -1.343114
         V35 |  -.5605039   .1499392    -3.74   0.001      -.86592   -.2550878
         V29 |  -2.794011   .9732333    -2.87   0.007    -4.776423   -.8116002
         V30 |  -1.453439   .2613666    -5.56   0.000    -1.985825   -.9210525
         V31 |   -1.23254   .4398462    -2.80   0.009    -2.128477   -.3366022
         V37 |   1.592627   .5664019     2.81   0.008     .4389036    2.746349
         V36 |   3.094153   .6995871     4.42   0.000     1.669141    4.519165
         V34 |   2.968051   .7713446     3.85   0.001     1.396873    4.539229
       _cons |  -27.89418   112.8038    -0.25   0.806     -257.668    201.8797
------------------------------------------------------------------------------
```

图 23.33　回归分析结果图 5

在上述分析结果中，我们可以得到很多信息。可以看出共有 42 个样本参与了分析，模型的 F 值(9, 32) =708.01，P 值（Prob > F）＝0.0000，说明模型整体上是非常显著的。模型的可决系数（R-squared）为 0.9950，模型修正的可决系数（Adj R-squared）为 0.9936，说明模型的解释能力是非常优秀接近完美的。

模型经过四次剔除变量后得到最终结果。第一个模型是包含全部自变量的全模型，该模型中 V26 变量的系数显著性 P 值高达 0.9855，被剔除掉；第二个模型是剔除掉自变量 V26 以后的模型，该模型中 V32 变量的系数显著性 P 值高达 0.9251，被剔除掉；第三个模型是剔除掉自变量 V26、V32 以后的模型，该模型中 V33 变量的系数显著性 P 值高达 0.4340，被剔除掉；第四个模型是剔除掉自变量 V26、V32、V33 以后的模型，该模型中 V28 变量的系数显著性 P 值高达 0.1746，被剔除掉。剔除掉自变量 V26、V32、V33、V28 以后，我们得到最终回归模型。

在最终回归模型中，变量 V38 的系数标准误是 0.8543404，t 值为 2.73，P 值为 0.010，系数是非常显著的，95%的置信区间为[0.5907202, 4.071189]。变量 V27 的系数标准误是 0.8334345，t 值为-3.65，P 值为 0.001，系数是非常显著的，95%的置信区间为[-4.738415, -1.343114]。变量 V35 的系数标准误是 0.1499392，t 值为-3.74，P 值为 0.001，系数是非常显著的，95%的置信区间为[-0.86592, -0.2550878]。变量 V29 的系数标准误是 0.9732333，t 值为-2.87，P 值为 0.007，系数是非常显著的，95%的置信区间为[-4.776423, -0.8116002]。变量 V30 的系数标准误是 0.2613666，t 值为-5.56，P 值为 0.000，系数是非常显著的，95%的置信区间为[-1.985825, -0.9210525]。变量 V31 的系数标准误是 0.4398462，t 值为-2.80，P 值为 0.009，系数是非常显著的，95%的置信区间为[-2.128477, -0.3366022]。变量 V37 的系数标准误是 0.5664019，t 值为 2.81，P 值为 0.008，系数是非常显著的，95%的置信区间为[0.4389036, 2.746349]。变量 V36 的系数标准误是.6995871，t 值为 4.42，P 值为 0.000，系数是非常显著的，95%的置信区间为[1.669141, 4.519165]。变量 V34 的系数标准误是 0.7713446，t 值为 3.85，P 值为 0.001，系数是非常显著的，95%的置信区间为[1.396873, 4.539229]。常数项的系数标准误是 112.8038，t 值为-0.25，P 值为 0.806，系数是非常不显著的，95%的置信区间为[-257.668,

201.8797]。

最终最小二乘回归模型的方程是：

赔款支出合计= -3.040764 *女-2.794011 *硕士-1.453439 *学士-1.23254 *大专+ 2.968051 *中级-0.5605039 *初级+ 3.094153 *三十五岁以下+ 1.592627 *三十六岁到四十五岁+ 2.330955 *四十六岁以上+ u

经过以上最小二乘回归分析，可以发现我国财产保险公司的总赔款支出水平与公司职员的性别、年龄、职称、文化水平都有一定的显著关系。具体而言，中级职称或者三十五岁以下、三十六岁到四十五岁、四十六岁以上的职员对公司的总赔款支出有拉动效应；硕士学历、学士学历、大专学历或者初级职称或者女性的职员对公司的总赔款支出有拖后效应。

（2）图 23.34 是以“赔款支出合计”为因变量，以“男”“女”“博士”“硕士”“学士”“大专”“中专以下”“高级”“中级”“初级”“三十五岁以下”“三十六岁到四十五岁”“四十六岁以上”为自变量，并使用以“company”为聚类变量的聚类稳健标准差，进行最小二乘回归分析的结果。

```
. reg  V14  V26-V38,vce(cluster  company)

Linear regression                               Number of obs =         42
                                                F(  8,     41) =          .
                                                Prob > F      =          .
                                                R-squared     =     0.9954
                                                Root MSE      =     619.01

                            (Std. Err. adjusted for 42 clusters in company)
------------------------------------------------------------------------------
             |               Robust
         V14 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         V26 |   -5.85334   147.8865    -0.04   0.969    -304.5162    292.8095
         V27 |  -7.622048   148.9048    -0.05   0.959    -308.3414    293.0974
         V28 |  -51.51143   197.0673    -0.26   0.795     -449.497    346.4742
         V29 |  -7.412917   145.8779    -0.05   0.960    -302.0192    287.1934
         V30 |  -6.574428   145.9988    -0.05   0.964     -301.425    288.2761
         V31 |  -6.517744   146.1778    -0.04   0.965    -301.7297    288.6943
         V32 |  -5.356681   146.1497    -0.04   0.971    -300.5119    289.7986
         V33 |  -4.732002   7.031998    -0.67   0.505    -18.93341    9.469406
         V34 |   4.187575    1.93891     2.16   0.037     .2718671    8.103282
         V35 |  -.4520437    .229336    -1.97   0.055     -.915197    .0111097
         V36 |    13.4855   5.938586     2.27   0.028      1.49228    25.47872
         V37 |   12.08312   5.730819     2.11   0.041     .5095007    23.65675
         V38 |   12.78781   5.596631     2.28   0.028      1.48519    24.09044
       _cons |   95.38777   172.7519     0.55   0.584    -253.4918    444.2673
------------------------------------------------------------------------------
```

图 23.34　回归分析结果图 6

我们可以看出，该结果中有很多变量系数的显著性是非常差的，需要把不显著的变量进行剔除后再进行进一步分析。

（3）图 23.35 是在上步回归的基础上，剔除掉不显著的自变量以后，以“赔款支出合计”为因变量，以 V26~V30、V33~V38 为自变量，并使用以“company”为聚类变量的聚类稳健标准差，进行最小二乘回归分析的结果。

```
. reg  V14  V27  V29-V31  V34-V38,vce(cluster  company) nocon

Linear regression                               Number of obs =        42
                                                F(  9,     41) =83669.89
                                                Prob > F      =   0.0000
                                                R-squared     =   0.9955
                                                Root MSE      =   594.47

                               (Std. Err. adjusted for 42 clusters in company)
------------------------------------------------------------------------------
             |               Robust
         V14 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         V27 |  -3.061834   1.251517    -2.45   0.019    -5.589323   -.5343447
         V29 |  -2.809717   1.312409    -2.14   0.038     -5.46018   -.1592543
         V30 |   -1.46242   .4131895    -3.54   0.001    -2.296873   -.6279665
         V31 |  -1.248829   .5566316    -2.24   0.030    -2.372969   -.1246884
         V34 |   2.998872   .9540455     3.14   0.003     1.072138    4.925606
         V35 |  -.5609798   .1958466    -2.86   0.007    -.9565002   -.1654595
         V36 |   3.116446   1.038834     3.00   0.005     1.018478    5.214415
         V37 |   1.600188   .7876351     2.03   0.049     .0095266    3.190849
         V38 |   2.326594   1.041013     2.23   0.031     .2242263    4.428962
------------------------------------------------------------------------------
```

图 23.35　回归分析结果图 7

可以看出，在剔除掉不显著的自变量以后，以“赔款支出合计”为因变量，以 V27、V29~V31、V34~V38 为自变量，并使用以“company”为聚类变量的聚类稳健标准差，进行最小二乘回归分析的结果与普通最小二乘回归分析有所区别。

在该模型中，最终保留的自变量有“女”“硕士”“学士”“大专”“中级”“初级”“三十五岁以下”“三十六岁到四十五岁”“四十六岁以上”。

该模型方程为：

赔款支出合计= -3.061834 *女-2.809717*硕士-1.46242 *学士-1.248829 *大专+2.998872 *中级-0.5609798*初级+3.116446*三十五岁以下+ 1.600188 *三十六岁到四十五岁+ 2.326594 *四十六岁以上+ u

经过以上分析，可以发现我国财产保险公司的总赔款支出水平与公司职员的性别、年龄、职称、文化水平都有一定的显著关系。具体而言，中级职称或者三十五岁以下、三十六岁到四十五岁、四十六岁以上的职员对公司的总赔款支出有拉动效应；初级职称或者硕士学历、学士学历、大专学历或者女性的职员对公司的总赔款支出有拖后效应。

23.7　因子分析

对于因子分析，我们准备从以下两部分进行：

第一，对构成保费收入的各个变量提取公因子。

第二，对构成赔款支出的各个变量提取公因子。

1. 对构成保费收入的各个变量提取公因子

操作步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出“主界面”对话框。

02 在“主界面”对话框的“Command”文本框中分别输入下面的命令并按键盘上的回车键进行确认：

（1）factor V3-V11,pcf

本命令的含义是采用主成分因子法对变量 V3-V11 进行因子分析。

（2）rotate

本命令的含义是采用最大方差正交旋转法对因子结构进行旋转。

（3）loadingplot,factors(2) yline(0) xline(0)

本命令的含义是绘制因子旋转后的因子载荷图。

（4）predict f1 f2

本命令的含义是展示因子分析后各个样本的因子得分情况。

（5）correlate　f1 f2

本命令的含义是展示系统提取的两个主因子的相关系数矩阵。

（6）scoreplot,mlabel(V1) yline(0) xline(0)

本命令的含义是展示每个样本的因子得分示意图。

（7）estat kmo

本命令的含义是展示本例因子分析的 KMO 检验结果。

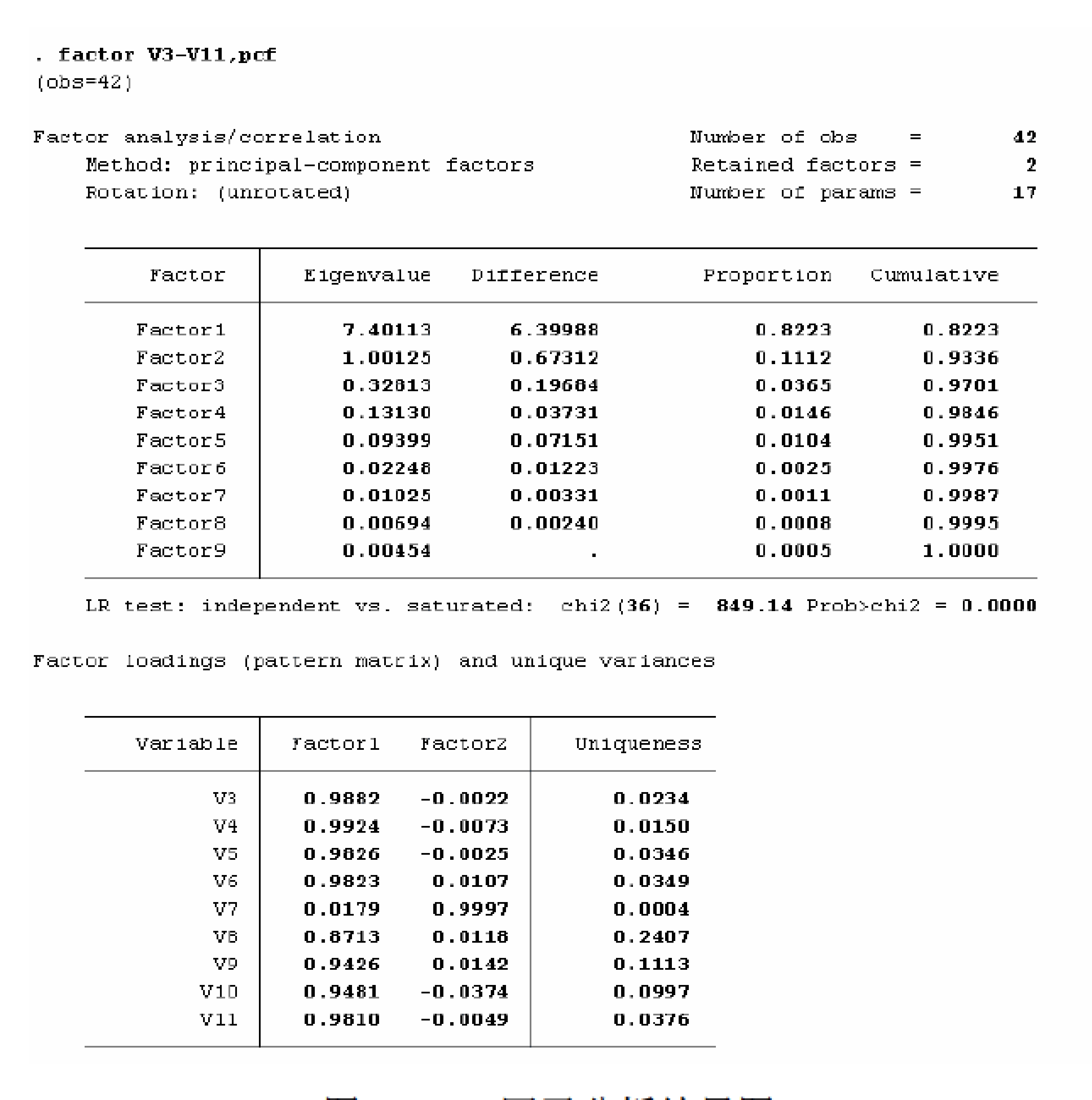
（8）screeplot

本命令的含义是展示本例因子分析所提取的各个因子的特征值碎石图。

03 设置完毕，等待输出结果。

在Stata 14.0“主界面”的结果窗口我们可以看到如图23.36~图23.50所示的分析结果。

（1）图23.36展示的是因子分析的基本情况。



```
. factor V3-V11,pcf
(obs=42)

Factor analysis/correlation                    Number of obs    =        42
    Method: principal-component factors        Retained factors =         2
    Rotation: (unrotated)                      Number of params =        17
```

Factor	Eigenvalue	Difference	Proportion	Cumulative
Factor1	7.40113	6.39988	0.8223	0.8223
Factor2	1.00125	0.67312	0.1112	0.9336
Factor3	0.32813	0.19604	0.0365	0.9701
Factor4	0.13130	0.03731	0.0146	0.9846
Factor5	0.09399	0.07151	0.0104	0.9951
Factor6	0.02248	0.01223	0.0025	0.9976
Factor7	0.01025	0.00331	0.0011	0.9987
Factor8	0.00694	0.00240	0.0008	0.9995
Factor9	0.00454	.	0.0005	1.0000

LR test: independent vs. saturated: chi2(36) = 849.14 Prob>chi2 = 0.0000

Factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Uniqueness
V3	0.9882	-0.0022	0.0234
V4	0.9924	-0.0073	0.0150
V5	0.9826	-0.0025	0.0346
V6	0.9823	0.0107	0.0349
V7	0.0179	0.9997	0.0004
V8	0.8713	0.0118	0.2407
V9	0.9426	0.0142	0.1113
V10	0.9481	-0.0374	0.0997
V11	0.9810	-0.0049	0.0376

图 23.36　因子分析结果图 1

图23.36的上半部分说的是因子分析模型的一般情况，从图中我们可以看出共有42个样本（Number of obs= 42）参与了分析，提取保留的因子共有两个（Retained factors = 2），模型LR检验的卡方值（LR test: independent vs. saturated: chi2(36)）为849.14，P值（Prob>chi2）为0.0000，模型非常显著。图23.36的上半部分最左列（Factor）说明的是因子名称，可以看出模型共提取了9个因子。Eigenvalue列表示的是提取因子的特征值情况，只有前两个因子的特征值是大于1的，其中第一个因子的特征值是7.40113，第二个因子的特征值是1.00125。Proportion列表示的是提取因子的方差贡献率，其中第一个因子的方差贡献率为82.23%，第二个因子的方差贡献率为11.12%。Cumulative列表示的是提取因子的累计方差贡献率，其中前两个因子的累计方差贡献率为93.36%。

图23.36的下半部分说的是模型的因子载荷矩阵以及变量的未被解释部分。其中，Variable列表示的是变量名称，Factor1、Factor2两列分别说明的是提取的前两个主因子（特征值大于1的）对各个变量的解释程度，本例中，Factor1主要解释的是V3、V4、V5、V6、V8、V9、V10、V11这8个变量的信息，Factor2主要解释的是V7变量的信息。Uniqueness列表示变量未被提取的前两个主因子解释的部分，可以发现在舍弃其他主因子的情况下，信息的损失量是很小的。

（2）图23.37展示的是对因子结构进行旋转的结果。学者们的研究表明，旋转操作有助于进一步简化因子结构。Stata 14.0支持的旋转方式有两种，一种是最大方差正交旋转，一般适用于相互独立的因子或者成分，也是系统默认的情况；另外一种是promax斜交旋转，它允许因子或者成分之间存在相关关系。此处我们选择系统默认方式，当然我们后面的操作也证明了这样做的恰当性。

```
. rotate

Factor analysis/correlation                  Number of obs    =        42
    Method: principal-component factors      Retained factors =         2
    Rotation: orthogonal varimax (Kaiser off) Number of params =        17
```

Factor	Variance	Difference	Proportion	Cumulative
Factor1	7.40037	6.39837	0.8223	0.8223
Factor2	1.00201	.	0.1113	0.9336

LR test: independent vs. saturated: chi2(36) = 849.14 Prob>chi2 = 0.0000

Rotated factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Uniqueness
V3	0.9882	0.0086	0.0234
V4	0.9925	0.0035	0.0150
V5	0.9825	0.0082	0.0346
V6	0.9822	0.0213	0.0349
V7	0.0071	0.9998	0.0004
V8	0.8711	0.0213	0.2407
V9	0.9424	0.0244	0.1113
V10	0.9485	-0.0271	0.0997
V11	0.9810	0.0058	0.0376

Factor rotation matrix

	Factor1	Factor2
Factor1	0.9999	0.0109
Factor2	-0.0109	0.9999

图 23.37　因子分析结果图 2

图23.37包括3部分内容，第一部分说的是因子旋转模型的一般情况，从图中我们可以看出共有42个样本（Number of obs = 42）参与了分析，提取保留的因子共有两个（Retained factors = 2），模型LR检验的卡方值（LR test: independent vs. saturated: chi2(15)）为849.14，P值（Prob>chi2）为0.0000，模型非常显著。最左列（Factor）说明的是因子名称，可以看出模型旋转后共提取了2个因子。Proportion列表示的是提取因子的方差贡献率，其中第一个因子的方差贡献率为82.23%，第二个因子的方差贡献率为11.13%。Cumulative列表示的是提取因子的累计方差贡献率，其中前两个因子的累计方差贡献率为93.36%。

图23.37的第二部分说的是模型的因子载荷矩阵以及变量的未被解释部分。其中Variable列表示的是变量名称，Factor1、Factor2两列分别说明的是旋转提取的两个主因子对各个变量的解释程度，本例中，Factor1主要解释的是V3、V4、V5、V6、V8、V9、V10、V11这8个变量的信息，Factor2主要解释的是V7变量的信息。Uniqueness列表示变量未被提取的前两个主因子解释的部分，可以发现在舍弃其他主因子的情况下，信息的损失量是很小的。

图23.37的第三部分展示的是因子旋转矩阵的一般情况，提取的两个因子不存在相关关系。

（3）图23.38展示的是因子旋转后的因子载荷图。因子载荷图可以使用户更加直观地看出各个变量被两个因子解释的情况。

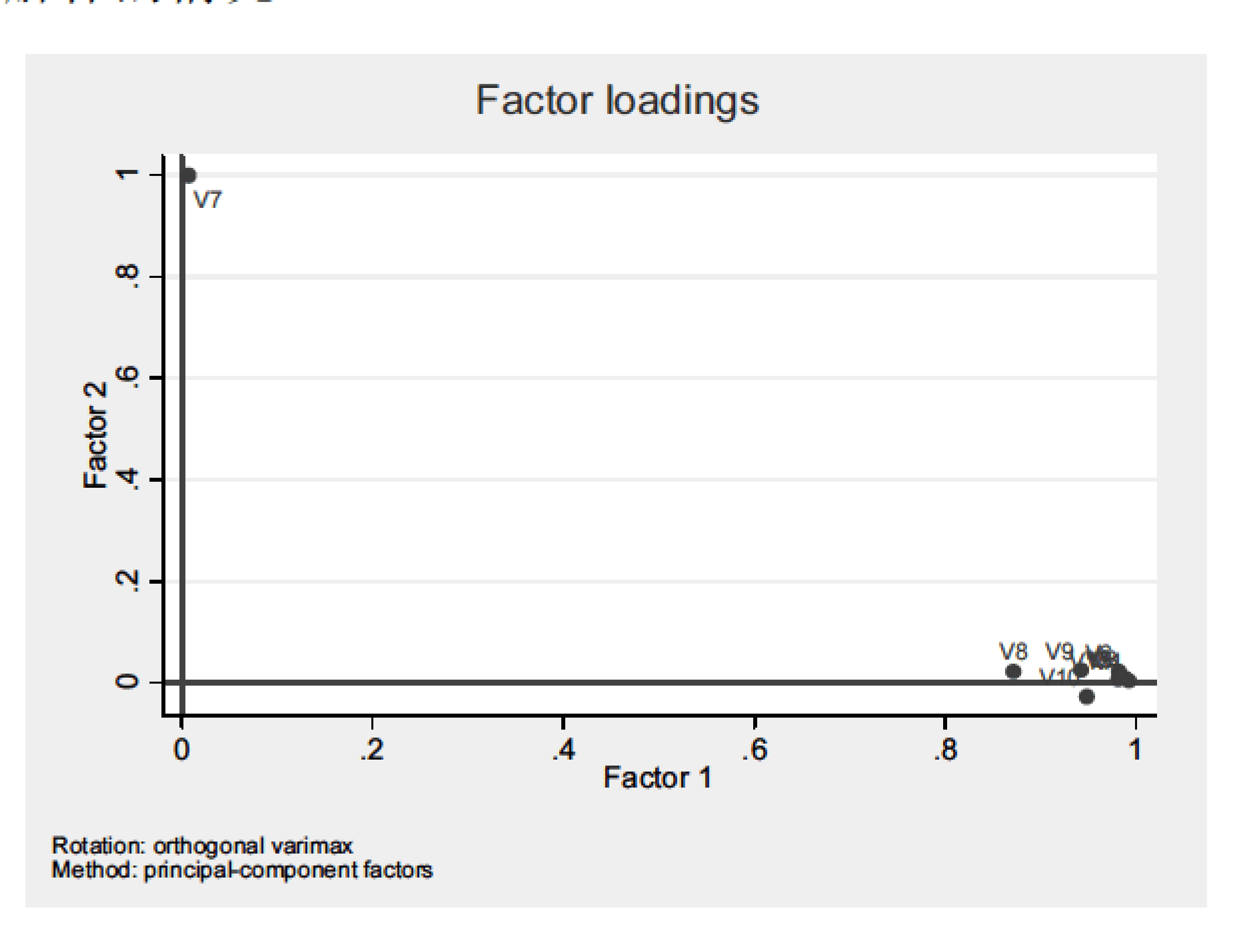


图 23.38　因子分析结果图 3

与前面的分析相同，我们发现V3、V4、V5、V6、V8、V9、V10、V11这8个变量的信息主要被Factor1这一因子所解释，V7变量主要被Factor2这一因子所解释。

（4）图23.39展示的是因子分析后各个样本的因子得分情况。因子得分的概念是通过将每个变量标准化为平均数等于0和方差等于1，然后以因子分析系数进行加权合计为每个因子构成的线性情况。以因子的方差贡献率为权数对因子进行加权求和，即可得到每个样本的因子综合得分。

```
. predict f1 f2
(regression scoring assumed)

Scoring coefficients (method = regression; based on varimax rotated factors)
```

Variable	Factor1	Factor2
V3	0.13354	-0.00073
V4	0.13416	-0.00582
V5	0.13278	-0.00105
V6	0.13260	0.01209
V7	-0.00843	0.99837
V8	0.11759	0.01311
V9	0.12719	0.01555
V10	0.12850	-0.03598
V11	0.13259	-0.00341

图 23.39 因子分析结果图 4

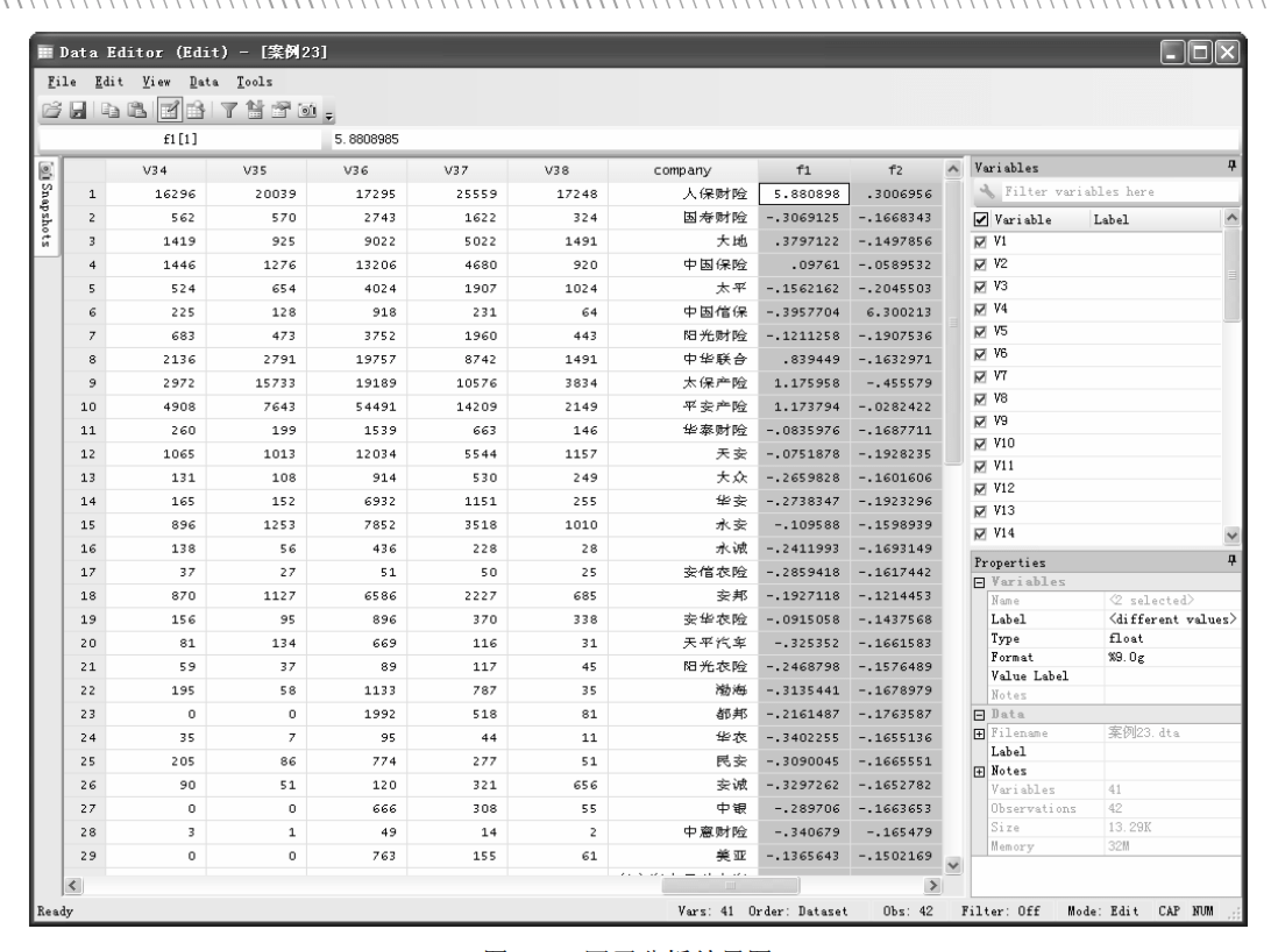
根据图 23.39 展示的因子得分系数矩阵，我们可以写出各公因子的表达式。值得一提的是，在表达式中各个变量已经不是原始变量，而是标准化变量。

表达式如下：

F1=0.134*企业财产保险保费收入+0.134*机动车辆保险保费收入
+0.133*货物运输保险保费收入+0.133*责任保险保费收入
−0.008*信用保证保险保费收入+0.118*农业保险保费收入
+0.127*短期健康保险保费收入+0.128*意外伤害保险保费收入
+0.133*其他保险保费收入

F2=0.000*企业财产保险保费收入−0.006*机动车辆保险保费收入
−0.001*货物运输保险保费收入+0.012*责任保险保费收入
+0.998*信用保证保险保费收入+0.013*农业保险保费收入
+0.015*短期健康保险保费收入−0.036*意外伤害保险保费收入
−0.004*其他保险保费收入

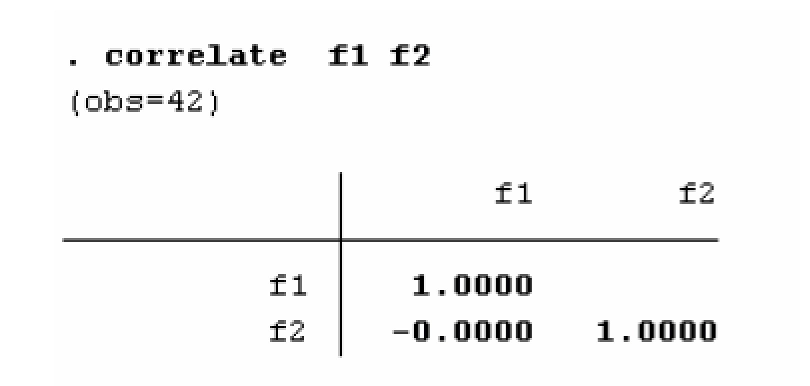
选择“Data” | “Data Editor” | “Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 23.40 所示的因子得分数据。



	V34	V35	V36	V37	V38	company	f1	f2
1	16296	20039	17295	25559	17248	人保财险	5.880898	.3006956
2	562	570	2743	1622	324	国寿财险	-.3069125	-.1668343
3	1419	925	9022	5022	1491	大地	.3797122	-.1497856
4	1446	1276	13206	4680	920	中国保险	.09761	-.0589532
5	524	654	4024	1907	1024	太平	-.1562162	-.2045503
6	225	128	918	231	64	中国信保	-.3957704	6.300213
7	683	473	3752	1960	443	阳光财险	-.1211258	-.1907536
8	2136	2791	19757	8742	1491	中华联合	.839449	-.1632971
9	2972	15733	19189	10576	3834	太保产险	1.175958	-.455579
10	4908	7643	54491	14209	2149	平安产险	1.173794	-.0282422
11	260	199	1539	663	146	华泰财险	-.0835976	-.1687711
12	1065	1013	12034	5544	1157	天安	-.0751878	-.1928235
13	131	108	914	530	249	大众	-.2659828	-.1601606
14	165	152	6932	1151	255	华安	-.2738347	-.1923296
15	896	1253	7852	3518	1010	永安	-.109588	-.1598939
16	138	56	436	228	28	永诚	-.2411993	-.1693149
17	37	27	51	50	25	安信农险	-.2859418	-.1617442
18	870	1127	6586	2227	685	安邦	-.1927118	-.1214453
19	156	95	896	370	338	安华农险	-.0915058	-.1437568
20	81	134	669	116	31	天平汽车	-.325352	-.1661583
21	59	37	89	117	45	阳光农险	-.2468798	-.1576489
22	195	58	1133	787	35	渤海	-.3135441	-.1678979
23	0	0	1992	518	81	都邦	-.2161487	-.1763587
24	35	7	95	44	11	华农	-.3402255	-.1655136
25	205	86	774	277	51	民安	-.3090045	-.1665551
26	90	51	120	321	656	安诚	-.3297262	-.1652782
27	0	0	666	308	55	中银	-.289706	-.1663653
28	3	1	49	14	2	中意财险	-.340679	-.165479
29	0	0	763	155	61	美亚	-.1365643	-.1502169

图 23.40 因子分析结果图 5

（5）图23.41展示的是系统提取的两个主因子的相关系数矩阵。



```
. correlate  f1 f2
(obs=42)

             |       f1       f2
-------------+------------------
          f1 |   1.0000
          f2 |  -0.0000   1.0000
```

图 23.41 因子分析结果图 6

从图23.41中可以看出，我们提取的两个主因子之间几乎没有什么相关关系，这也说明了我们在前面对因子进行旋转的操作环节中采用最大方差正交旋转方式是明智的。值得说明的是图中f1与f2的相关系数是-0.0000并非是不正确的，这是由于Stata 14.0只保留了4位小数所导致的，比如真实的数据有可能是-0.00001，那么结果显示的就是-0.0000。

（6）图23.42展示的是每个样本的因子得分示意图。

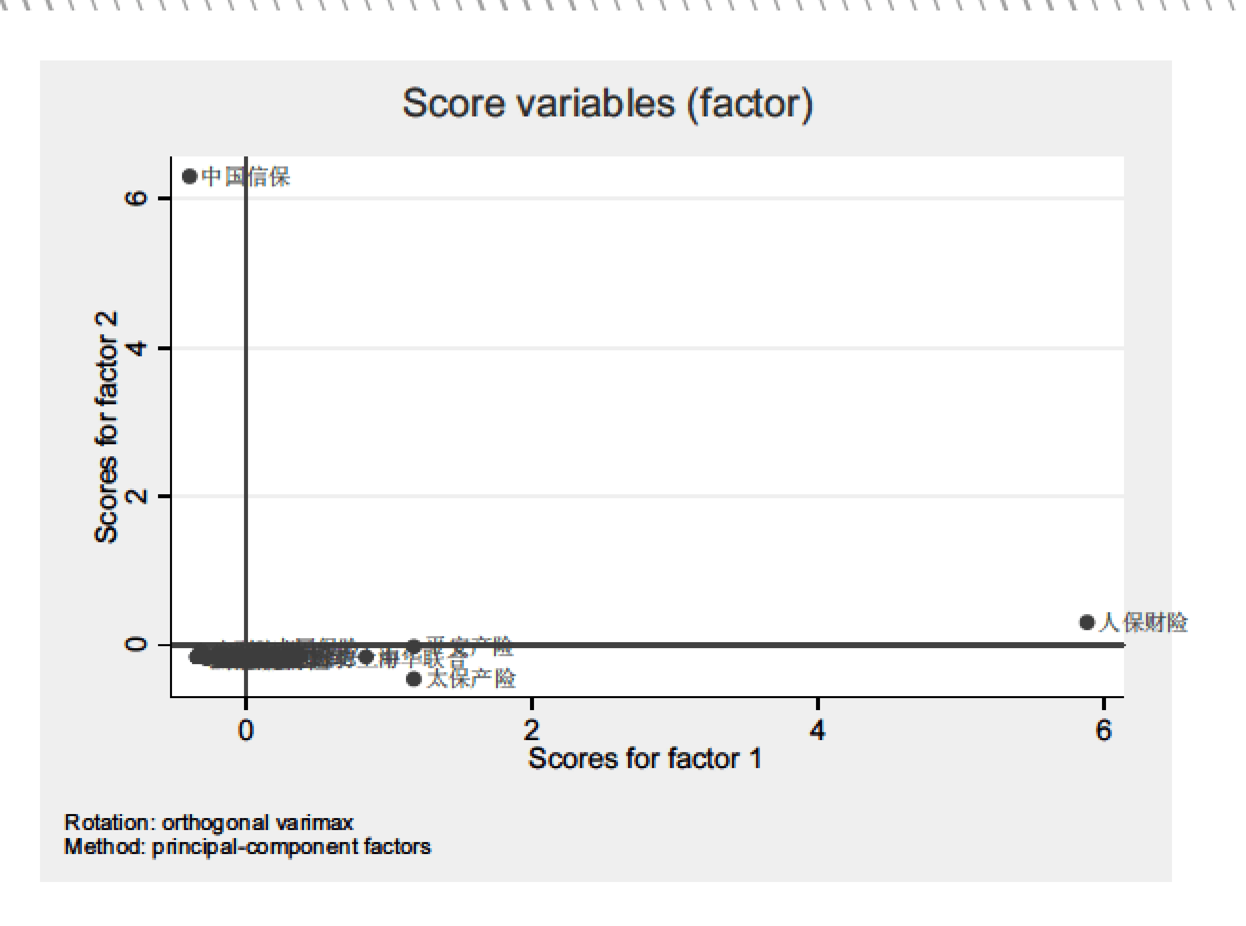


图 23.42 因子分析结果图 7

从图23.42中可以看出，所有的样本被分到四个象限，我们可以比较直观地看出各个样本的因子得分分布情况。

（7）图23.43展示的是本例因子分析的KMO检验结果。

```
. estat kmo

Kaiser-Meyer-Olkin measure of sampling adequacy

    -----------------------
        Variable |     kmo
    -------------+---------
              V3 |  0.9039
              V4 |  0.8543
              V5 |  0.9138
              V6 |  0.9122
              V7 |  0.2315
              V8 |  0.9268
              V9 |  0.9121
             V10 |  0.8746
             V11 |  0.9036
    -------------+---------
         Overall |  0.8986
    -----------------------
```

图 23.43 因子分析结果图 8

KMO 检验是为了看数据是否适合进行因子分析，其取值范围是 0~1。其中，0.9~1 表示极好，0.8~0.9 表示可奖励的，0.7~0.8 表示还好，0.6~0.7 表示中等，0.5~0.6 表示糟糕，0~0.5 表示不可接受。如图 23.43 所示，本例中总体（Overall）KMO 的取值为 0.8986，表明可以进行因子分析。各个变量的 KMO 值也大多在 0.8 以上，所以本例是比较适合因子分析的，模型的构建是有意义的。

（8）图23.44展示的是本例因子分析所提取的各个因子的特征值碎石图。

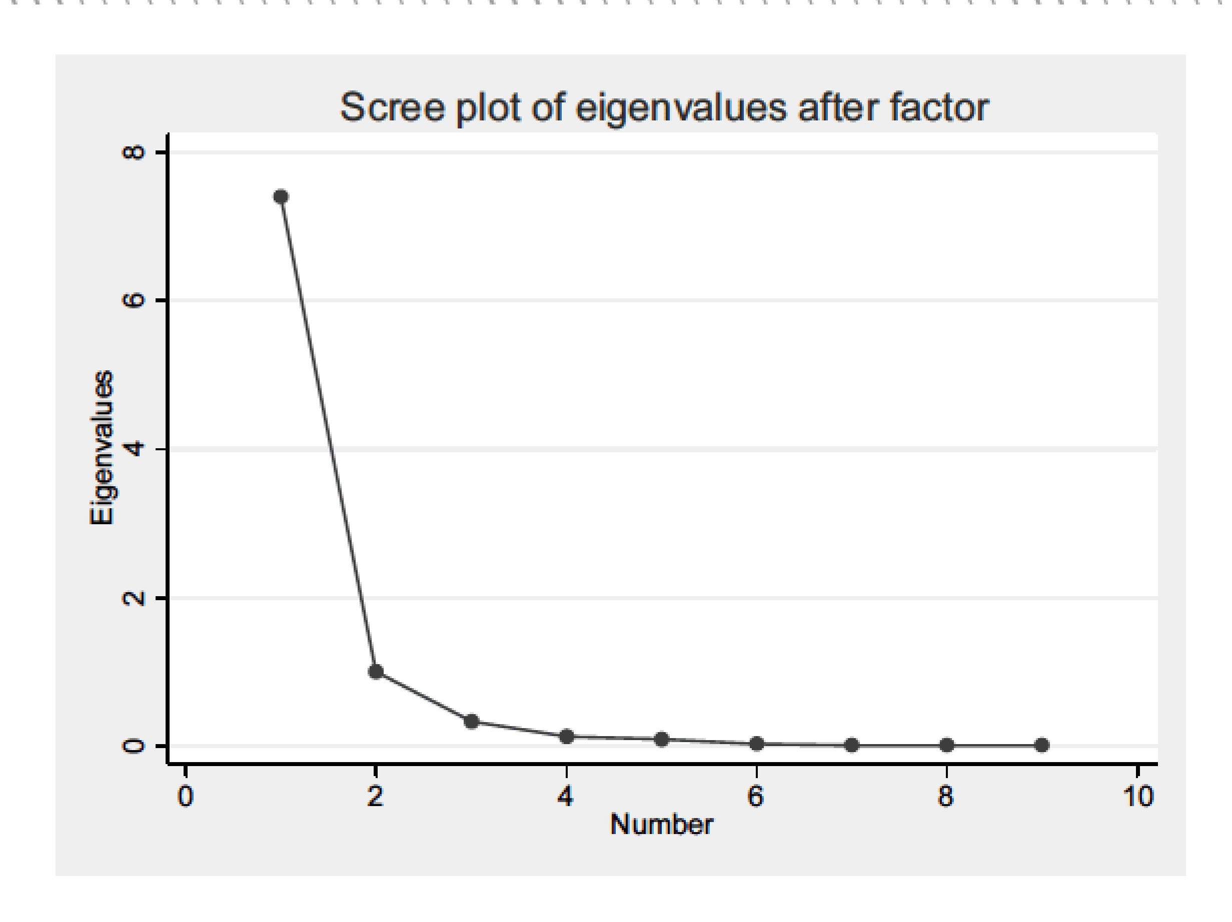


图 23.44 因子分析结果图 9

碎石图可以非常直观地观测出提取因子的特征值大小情况。图 23.44 的横轴表示的是系统提取因子的名称，并且已经按特征值大小进行降序排列好，纵轴表示因子特征值的大小情况。从图 23.44 中可以轻松地看出本例中只有前两个因子的特征值是大于 1 的。

2. 对构成赔款支出的各个变量提取公因子

操作步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出“主界面”对话框。

02 在“主界面”对话框的“Command”文本框中分别输入下面的命令并按键盘上的回车键进行确认：

（1）factor V15-V23,pcf

本命令的含义是采用主成分因子法对变量 V15-V23 进行因子分析。

（2）rotate

本命令的含义是采用最大方差正交旋转法对因子结构进行旋转。

（3）predict f1

本命令的含义是展示因子分析后各个样本的因子得分情况。

（4）estat kmo

本命令的含义是展示本例因子分析的 KMO 检验结果。

（5）screeplot

本命令的含义是展示本例因子分析所提取的各个因子的特征值碎石图。

03 设置完毕，等待输出结果。

在 Stata 14.0“主界面”的结果窗口我们可以看到如图 23.45~图 23.50 所示的分析结果。

（1）图 23.45 展示的是因子分析的基本情况。

```
. factor V15-V23,pcf
(obs=42)

Factor analysis/correlation                        Number of obs    =        42
    Method: principal-component factors            Retained factors =         1
    Rotation: (unrotated)                          Number of params =         9

    --------------------------------------------------------------------------
         Factor  |   Eigenvalue   Difference        Proportion   Cumulative
    -------------+------------------------------------------------------------
        Factor1  |      7.28382      6.56972            0.8093       0.8093
        Factor2  |      0.71410      0.11070            0.0793       0.8887
        Factor3  |      0.60341      0.37245            0.0670       0.9557
        Factor4  |      0.23095      0.11247            0.0257       0.9814
        Factor5  |      0.11848      0.08619            0.0132       0.9945
        Factor6  |      0.03230      0.02487            0.0036       0.9981
        Factor7  |      0.00742      0.00094            0.0008       0.9989
        Factor8  |      0.00648      0.00346            0.0007       0.9997
        Factor9  |      0.00303            .            0.0003       1.0000
    --------------------------------------------------------------------------
    LR test: independent vs. saturated:  chi2(36) =  825.54 Prob>chi2 = 0.0000

Factor loadings (pattern matrix) and unique variances

    ---------------------------------------
        Variable |  Factor1 |   Uniqueness
    -------------+----------+--------------
             V15 |   0.9850 |      0.0298
             V16 |   0.9834 |      0.0329
             V17 |   0.9749 |      0.0495
             V18 |   0.9736 |      0.0522
             V19 |   0.6262 |      0.6079
             V20 |   0.6223 |      0.6127
             V21 |   0.9549 |      0.0882
             V22 |   0.8912 |      0.2057
             V23 |   0.9812 |      0.0373
    ---------------------------------------
```

图 23.45 因子分析结果图 10

图 23.45 的上半部分说的是因子分析模型的一般情况，从图中我们可以看出共有 42 个样本（Number of obs= 42）参与了分析，提取保留的因子共有 1 个（Retained factors = 1），模型 LR 检验的卡方值（LR test: independent vs. saturated: chi2(36)）为 825.54，P 值（Prob>chi2）为 0.0000，模型非常显著。图 23.45 的上半部分最左列（Factor）说明的是因子名称，可以看出模型共提取了 9 个因子。Eigenvalue 列表示的是提取因子的特征值情况，只有第一个因子的特征值是大于 1 的，第一个因子的特征值是 7.28382。Proportion 列表示的是提取因子的方差贡献率，其中第一个因子的方差贡献率为 80.93%。Cumulative 列表示的是提取因子的累计方差贡献率，其中前两个因子的累计方差贡献率为 88.87%。

图 23.45 的下半部分说的是模型的因子载荷矩阵以及变量的未被解释部分。其中，Variable 列表示的是变量名称，Factor1 列说明的是提取的第一个主因子（特征值大于 1 的）对各个变量的解释程度。Uniqueness 列表示变量未被提取的第一主因子解释的部分，可以发现在舍弃其他主因子的情况下，信息的损失量是很小的。

（2）图 23.46 展示的是对因子结构进行旋转的结果。学者们的研究表明，旋转操作有助于进一步简化因子结构。Stata 14.0 支持的旋转方式有两种，一种是最大方差正交旋转，一般适用于相互独立的因子或者成分，也是系统默认的情况；另一种是 promax 斜交旋转，允许因子或者成分之间存在相关关系。此处我们选择系统默认方式，当然我们后面的操作也证明了这样做的恰当性。

```
. rotate

Factor analysis/correlation                        Number of obs    =        42
    Method: principal-component factors           Retained factors =         1
    Rotation: orthogonal varimax (Kaiser off)      Number of params =         9

    --------------------------------------------------------------------------
         Factor  |     Variance   Difference        Proportion   Cumulative
    -------------+------------------------------------------------------------
        Factor1  |      7.28382            .            0.8093       0.8093
    --------------------------------------------------------------------------
    LR test: independent vs. saturated:  chi2(36) =  825.54 Prob>chi2 = 0.0000

Rotated factor loadings (pattern matrix) and unique variances

    -------------------------------------
        Variable |  Factor1 | Uniqueness
    -------------+----------+------------
             V15 |   0.9850 |     0.0298
             V16 |   0.9834 |     0.0329
             V17 |   0.9749 |     0.0495
             V18 |   0.9736 |     0.0522
             V19 |   0.6262 |     0.6079
             V20 |   0.6223 |     0.6127
             V21 |   0.9549 |     0.0882
             V22 |   0.8912 |     0.2057
             V23 |   0.9812 |     0.0373
    -------------------------------------

Factor rotation matrix

    --------------------------
                 |  Factor1
    -------------+------------
         Factor1 |   1.0000
    --------------------------
```

图 23.46 因子分析结果图 11

图23.46包括3部分内容，第一部分说的是因子旋转模型的一般情况，从图中我们可以看出共有42个样本（Number of obs = 42）参与了分析，提取保留的因子共有1个（Retained factors = 1），模型LR检验的卡方值（LR test: independent vs. saturated: chi2(15)）为825.54，P值（Prob>chi2）为0.0000，模型非常显著。最左列（Factor）说明的是因子名称，可以看出模型旋转后共提取了1个因子。Proportion列表示的是提取因子的方差贡献率，其中第一个因子的方差贡献率为80.93%。Cumulative列表示的是提取因子的累计方差贡献率。

图23.46的第二部分说的是模型的因子载荷矩阵以及变量的未被解释部分。其中，Variable列表示的是变量名称，Factor1列说明的是旋转提取的两个主因子对各个变量的解释程度，本例中，Factor1主要解释的是V15~V23这9个变量的信息， Uniqueness列表示变量未被提取的前两个主因子解释的部分，可以发现在舍弃其他主因子的情况下，信息的损失量是很小的。

图23.46的第三部分展示的是因子旋转矩阵的一般情况。

（3）图23.47展示的是因子分析后各个样本的因子得分情况。因子得分的概念是通过将每个变量标准化为平均数等于0和方差等于1，然后以因子分析系数进行加权合计为每个因子构成的线性情况。以因子的方差贡献率为权数对因子进行加权求和，即可得到每个样本的因子综合得分。

```
. predict f1
(regression scoring assumed)

Scoring coefficients (method = regression; based on varimax rotated factors)
```

Variable	Factor1
V15	0.13523
V16	0.13501
V17	0.13385
V18	0.13366
V19	0.08597
V20	0.08544
V21	0.13110
V22	0.12236
V23	0.13471

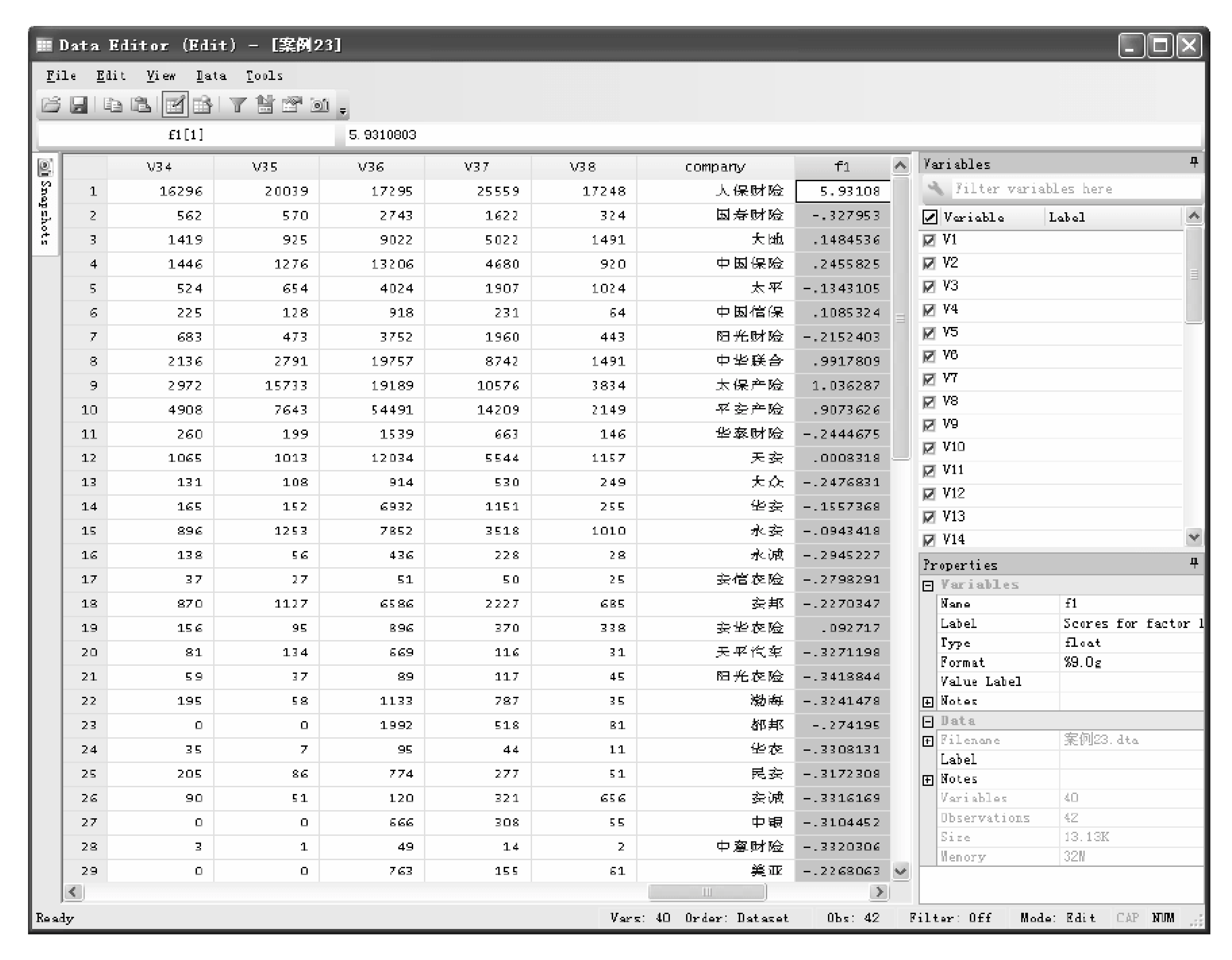
图 23.47 因子分析结果图 12

根据图 23.47 展示的因子得分系数矩阵，我们可以写出各公因子的表达式。值得一提的是，在表达式中各个变量已经不是原始变量而是标准化变量。

表达式如下：

F=0.135*企业财产保险赔款支出+0.135*机动车辆保险赔款支出
　+0.134 货物运输保险赔款支出+0.134*责任保险赔款支出
　+0.086*信用保证保险赔款支出+0.085*农业保险赔款支出
　+0.131*短期健康保险赔款支出+0.122*意外伤害保险赔款支出
　+0.135*其他保险赔款支出

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图 23.48 所示的因子得分数据。



Data Editor (Edit) – [案例23]

f1[1]　5.9310803

	V34	V35	V36	V37	V38	company	f1
1	16296	20039	17295	25559	17248	人保财险	5.93108
2	562	570	2743	1622	324	国寿财险	-.327953
3	1419	925	9022	5022	1491	大地	.1484536
4	1446	1276	13206	4680	920	中国保险	.2455825
5	524	654	4024	1907	1024	太平	-.1343105
6	225	128	918	231	64	中国信保	.1085324
7	683	473	3752	1960	443	阳光财险	-.2152403
8	2136	2791	19757	8742	1491	中华联合	.9917809
9	2972	15733	19189	10576	3834	太保产险	1.036287
10	4908	7643	54491	14209	2149	平安产险	.9073626
11	260	199	1539	663	146	华泰财险	-.2444675
12	1065	1013	12034	5544	1157	天安	.0008318
13	131	108	914	530	249	大众	-.2476831
14	165	152	6932	1151	255	华安	-.1557368
15	896	1253	7852	3518	1010	永安	-.0943418
16	138	56	436	228	28	永诚	-.2945227
17	37	27	51	50	25	安信农险	-.2798291
18	870	1127	6586	2227	685	安邦	-.2270347
19	156	95	896	370	338	安华农险	.092717
20	81	134	669	116	31	天平汽车	-.3271198
21	59	37	89	117	45	阳光农险	-.3418844
22	195	58	1133	787	35	渤海	-.3241478
23	0	0	1992	518	81	都邦	-.274195
24	35	7	95	44	11	华农	-.3308131
25	205	86	774	277	51	民安	-.3172308
26	90	51	120	321	656	安诚	-.3316169
27	0	0	666	308	55	中银	-.3104452
28	3	1	49	14	2	中意财险	-.3320306
29	0	0	763	155	61	美亚	-.2268063

图 23.48 因子分析结果图 13

（4）图23.49展示的是本例因子分析的KMO检验结果。

```
. estat kmo

Kaiser-Meyer-Olkin measure of sampling adequacy
```

Variable	kmo
V15	0.8929
V16	0.8041
V17	0.8587
V18	0.8435
V19	0.9717
V20	0.9011
V21	0.8001
V22	0.8004
V23	0.8127
Overall	0.8396

图 23.49 因子分析结果图 14

KMO 检验是为了看数据是否适合进行因子分析，其取值范围是 0~1。其中，0.9~1 表示极好，0.8~0.9 表示可奖励的，0.7~0.8 表示还好，0.6~0.7 表示中等，0.5~0.6 表示糟糕，0~0.5 表示不可接受。如图 23.49 所示，本例中总体（Overall）KMO 的取值为 0.8396，表明可以进行因子分析。全部变量的 KMO 值也都在 0.8 以上，所以本例是比较适合因子分析的，模型的构建是有意义的。

（5）图23.50展示的是本例因子分析所提取的各个因子的特征值碎石图。

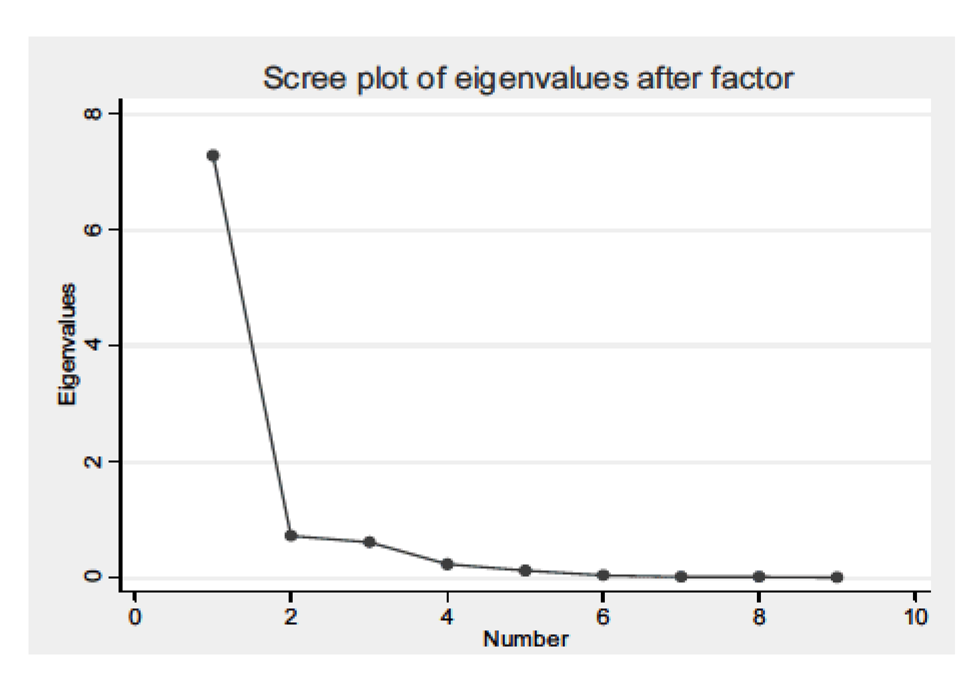


图 23.50 因子分析结果图 15

碎石图可以非常直观地观测出提取因子的特征值大小情况。图 23.50 的横轴表示的是系统提取因子的名称，并且已经按特征值大小进行降序排列好，纵轴表示因子特征值的大小情况。从图 23.50 中可以轻松地看出本例中只有第一个因子的特征值是大于 1 的。

23.8　聚类分析

对于聚类分析，我们也准备从两部分进行：

第一，使用构成保费收入的各个变量对各个财险公司进行聚类。

第二，使用构成赔款支出的各个变量对各个财险公司进行聚类。

1. 使用构成保费收入的各个变量对各个财险公司进行聚类

观察到不同变量的数量级相差不大，所以无须先对数据进行标准化处理，直接进行分析

即可。

分析步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出“主界面”对话框。

02 在“主界面”对话框的“Command”文本框中分别输入下面的命令并按键盘上的回车键进行确认：

cluster kmeans V3-V11,k(4)

本操作命令的含义是设定聚类数为 4，然后使用“K 个平均数的聚类分析”方法对变量 V3~V11 进行分析。

03 设置完毕，按键盘上的回车键，等待输出结果。

在 Stata 14.0“主界面”的结果窗口我们可以看到如图 23.51~图 23.54 所示的分析结果。

图23.51展示的是设定聚类数为4，然后使用“K个平均数的聚类分析”方法进行分析的结果。在输入Stata命令并分别按键盘上的回车键确认后，我们可以看到系统产生了一个新的变量，聚类变量_clus_1（cluster name: _clus_1）。

```
. cluster kmeans V3-V11,k(4)
cluster name: _clus_1
```

图 23.51　聚类分析结果图 1

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图23.52所示的_clus_1数据。

Data Editor (Edit) – [案例23]

_clus_1[1]　3

	V35	V36	V37	V38	company	f1	_clus_1
1	20039	17295	25559	17248	人保财险	5.93108	3
2	570	2743	1622	324	国寿财险	-.327953	2
3	925	9022	5022	1491	大地	.1484536	4
4	1276	13206	4680	920	中国保险	.2455825	4
5	654	4024	1907	1024	太平	-.1343105	4
6	128	918	231	64	中国信保	.1085324	2
7	473	3752	1960	443	阳光财险	-.2152403	4
8	2791	19757	8742	1491	中华联合	.9917809	1
9	15733	19189	10576	3834	太保产险	1.036287	1
10	7643	54491	14209	2149	平安产险	.9073626	1
11	199	1539	663	146	华泰财险	-.2444675	2
12	1013	12034	5544	1157	天安	.0008318	4
13	108	914	530	249	大众	-.2476831	2
14	152	6932	1151	255	华安	-.1557368	2
15	1253	7852	3518	1010	永安	-.0943418	4
16	56	436	228	28	永诚	-.2945227	2
17	27	51	50	25	安信农险	-.2798291	2
18	1127	6586	2227	685	安邦	-.2270347	4
19	95	896	370	338	安华农险	.092717	2
20	134	669	116	31	天平汽车	-.3271198	2
21	37	89	117	45	阳光农险	-.3418844	2
22	58	1133	787	35	渤海	-.3241478	2
23	0	1992	518	81	都邦	-.274195	2
24	7	95	44	11	华农	-.3308131	2
25	86	774	277	51	民安	-.3172308	2
26	51	120	321	656	安诚	-.3316169	2
27	0	666	308	55	中银	-.3104452	2
28	1	49	14	2	中意财险	-.3320306	2
29	0	763	155	61	美亚	-.2268063	2

图 23.52 聚类分析结果图 2

在图 23.52 中，我们可以看到所有的观测样本被分为四类，其中人保财险属于第三类，中

华联合、太保产险、平安产险属于第一类，大地、中国保险、太平、阳光财险、天安、永安、安邦属于第四类，其他财险公司属于第二类。可以发现，第三类公司各类保险的保费收入都非常高；第二类的信用保证保险保费收入较高，其他保险保费收入都很低；第一类信用保证保险保费收入很低，其他保险保费收入都较高；第四类的保险保费收入都较低，农业保险保费收入则很低。

我们通过聚类分析得到的研究结论是：人保财险各类保险的保费收入都非常高，是我国财产保险行业的“巨无霸”；中华联合、太保产险、平安产险信用保证保险保费收入很低，其他保险保费收入都较高；大地、中国保险、太平、阳光财险、天安、永安、安邦的保险收入则较低，农业保险保费收入很低；其他大部分的财险公司都是信用保证保险保费收入较高，而别的险种保费收入都很低，机动车辆保险保费收入和信用保证保险保费收入是其保费收入的最大来源。

2. 使用构成赔款支出的各个变量对各个财险公司进行聚类

观察到不同变量的数量级相差不大，所以无须先对数据进行标准化处理，直接进行分析即可。

分析步骤如下：

01 进入 Stata 14.0，打开相关数据文件，弹出“主界面”对话框。

02 在“主界面”对话框的“Command”文本框中分别输入下面的命令并按键盘上的回车键进行确认：

```
cluster kmeans V15-V23,k(4)
```

本操作命令的含义是设定聚类数为 4，然后使用“K 个平均数的聚类分析”方法对变量V15~V23 进行分析。

03 设置完毕，按键盘上的回车键，等待输出结果。

在Stata 14.0“主界面”的结果窗口我们可以看到如图23.53、图23.54所示的分析结果。

图23.53展示的是设定聚类数为4，然后使用“K个平均数的聚类分析”方法进行分析的结果。在输入Stata命令并分别按键盘上的回车键确认后，我们可以看到系统产生了一个新的变量，聚类变量_clus_2（cluster name: _clus_2）。

```
. cluster kmeans V15-V23,k(4)
cluster name: _clus_2
```

图 23.53　聚类分析结果图 3

选择“Data”|“Data Editor”|“Data Editor(Browse)”命令，进入数据查看界面，可以看到如图23.54所示的_clus_2数据。

Data Editor (Edit) - [案例23]

_clus_2[1]　4

	V36	V37	V38	company	f1	_clus_1	_clus_2
1	17295	25559	17248	人保财险	5.93108	3	4
2	2743	1622	324	国寿财险	-.327953	2	3
3	9022	5022	1491	大地	.1484536	4	2
4	13206	4680	920	中国保险	.2455825	4	3
5	4024	1907	1024	太平	-.1343105	4	3
6	918	231	64	中国信保	.1085324	2	3
7	3752	1960	443	阳光财险	-.2152403	4	3
8	19757	8742	1491	中华联合	.9917809	1	1
9	19189	10576	3834	太保产险	1.036287	1	1
10	54491	14209	2149	平安产险	.9073626	1	1
11	1539	663	146	华泰财险	-.2444675	2	3
12	12034	5544	1157	天安	.0008318	4	2
13	914	530	249	大众	-.2476831	2	3
14	6932	1151	255	华安	-.1557368	2	3
15	7852	3518	1010	永安	-.0943418	4	2
16	436	228	28	永诚	-.2945227	2	3
17	51	50	25	安信农险	-.2798291	2	3
18	6586	2227	685	安邦	-.2270347	4	2
19	896	370	338	安华农险	.092717	2	3
20	669	116	31	天平汽车	-.3271198	2	3
21	89	117	45	阳光农险	-.3418844	2	3
22	1133	787	35	渤海	-.3241478	2	3
23	1992	518	81	都邦	-.274195	2	3
24	95	44	11	华农	-.3308131	2	3
25	774	277	51	民安	-.3172308	2	3
26	120	321	656	安诚	-.3316169	2	3
27	666	308	55	中银	-.3104452	2	3
28	49	14	2	中意财险	-.3320306	2	3
29	763	155	61	美亚	-.2268063	2	3

Ready　Vars: 42　Order: Dataset　Obs: 42　Filter: Off　Mode: Edit

图 23.54 聚类分析结果图 4

图 23.54 中，我们可以看到所有的观测样本被分为四类，其中人保财险属于第四类，大地、天安、永安、安邦属于第二类，中华联合、太保产险、平安产险属于第一类，其他财险公司属于第三类。可以发现，第四类公司各类保险的赔款支出都非常高；第二类公司除信用保证保险赔款支出、农业保险赔款支出较低外，其他保险保费收入都最低；第三类公司则除信用保证保险赔款支出、农业保险赔款支出最低外，其他保险保费收入都较低；第一类各类保险的赔款支出都较高。

我们通过聚类分析得到的研究结论是：人保财险各类保险的赔款支出都非常高；大地、天安、永安、安邦等除信用保证保险赔款支出、农业保险赔款支出最低外，其他保险保费收入都较低；中华联合、太保产险、平安产险等各类保险的赔款支出都较高；其余财险公司除信用保证保险赔款支出、农业保险赔款支出较低外，其他保险赔款支出都最低。

23.9　研究结论

根据以上所做的分析，我们可以比较有把握地得出以下结论。

（1）简单相关分析表明：构成“保费收入合计”的 9 个组成部分，除“信用保证保险保费收入”与别的变量相关关系较弱外，其他变量之间都具有很强的相关性，都在 0.01 的显著性水平上显著。

（2）简单相关分析表明：构成“赔款支出合计”的所有变量之间都具有比较强的相关性，大部分的相关性还很强，在 0.01 的显著性水平上显著。

（3）简单相关分析表明：我国财险公司的“保费收入合计”“赔款支出合计”“总人数”这3个变量之间相关性很强。

（4）简单相关分析表明：我国财险公司的“赔案件数”“赔款支出合计”“未决赔款”这3个变量之间相关性很强。

（5）经过多重线性回归分析，可以发现我国财产保险公司的总保费收入水平与公司职员的性别、年龄、职称、文化水平都有一定的显著关系。具体而言，中级职称或者大专、中专以下、博士学历或者三十五岁以下、四十六岁以上的职员对公司的总保费收入有拉动效应，尤其是博士学历的职员，每增加一单位会带来对应保费收入的300多倍的增加；高级职称或者男性、女性的职员对公司的总保费收入有拖后效应。

（6）经过多重线性回归分析，可以发现我国财产保险公司的赔款支出总水平与公司职员的性别、年龄、职称、文化水平都有一定的显著关系。具体而言，中级职称或者三十五岁以下、三十六岁到四十五岁、四十六岁以上的职员对公司的总赔款支出有拉动效应；初级职称或者硕士学历、学士学历、大专学历或者女性的职员对公司的总赔款支出有拖后效应。

（7）因子分析表明：可以对构成我国财险公司“保费收入合计”的9个组成部分提取两个公因子，其中一个公因子主要反映除信用保证保险保费收入以外的变量的信息，第二个公因子反映的是信用保证保险保费收入这一变量的信息。

（8）因子分析表明：基于变量之间的高相关性，对构成我国财险公司“赔款支出合计”的9个组成部分提取一个公因子已足以反映这些变量的信息。

（9）聚类分析表明：人保财险各类保险的保费收入都非常高，是我国财产保险行业的“巨无霸”；太保产险、平安产险、华泰财险信用保证保险保费收入很低，其他保险保费收入都较高；大地、中国保险、太平、阳光财险、天安、永安、安邦农业保险保费收入很低，其他保险保费收入较低；剩余的大部分财险公司都是信用保证保险保费收入较高，而别的险种保费收入都很低，机动车辆保险保费收入和信用保证保险保费收入是其保费收入的最大来源。

（10）聚类分析表明：人保财险各类保险的赔款支出都非常高；大地、天安、永安、安邦等除信用保证保险赔款支出、农业保险赔款支出最低外，其他保险保费收入都较低；中华联合、太保产险、平安产险等各类保险的赔款支出都较高；其余的财险公司除信用保证保险赔款支出、农业保险赔款支出较低外，其他保险赔款支出都最低。

经过以上研究，我们可以从一种宏观的视野上对我国的财险公司有一个比较全面的了解，这对于以后我国财险公司的发展有重要的借鉴和指导意义。比如根据回归分析部分的结论，为提高总保费收入水平，我国财产保险公司在招聘员工的时候应该注意多招一些中级职称或者大专、中专以下、博士学历或者三十五岁以下、四十六岁以上的职员，为降低总赔款支出水平，我国财产保险公司在招聘员工的时候应该注意多招一些初级职称或者硕士学历、学士学历、大专学历或者女性职员。再如，聚类分析表明，人保财险在中国一枝独秀，大部分财险公司无论是保费收入还是赔款支出都相差甚远，所以为使我国财险业能以一种更加健康的充满竞争的方式成长，政府有必要做一些努力，以改变这种情况。

23.10　本章习题

使用《中国保险年鉴 2007》上的《中国 2006 年各保险公司人员结构情况统计》和《中国 2006 年各财产保险公司业务统计》数据（数据已整理入 Stata 中），进行以下分析。

（1）相关分析

第一，对“保费收入合计”的 9 个组成部分——“企业财产保险保费收入”“机动车辆保险保费收入”“货物运输保险保费收入”“责任保险保费收入”“信用保证保险保费收入”“农业保险保费收入”“短期健康保险保费收入”“意外伤害保险保费收入”“其他保险保费收入”进行简单相关分析。

第二，对“赔款支出合计”的 9 个组成部分——“企业财产保险赔款支出”“机动车辆保险赔款支出”“货物运输保险赔款支出”“责任保险赔款支出”“信用保证保险赔款支出”“农业保险赔款支出”“短期健康保险赔款支出”“意外伤害保险赔款支出”“其他保险赔款支出”进行简单相关分析。

第三，对“保费收入合计”“赔款支出合计”“总人数”这 3 个变量进行简单相关分析。

第四，对“赔案件数”“赔款支出合计”“未决赔款”这 3 个变量进行简单相关分析。

（2）回归分析

第一，以“保费收入合计”为因变量，以“男”“女”“博士”“硕士”“学士”“大专”“中专以下”“高级”“中级”“初级”“三十五岁以下”“三十六岁到四十五岁”“四十六岁以上”为自变量，进行最小二乘线性回归。

第二，以“赔款支出合计”为因变量，以“男”“女”“博士”“硕士”“学士”“大专”“中专以下”“高级”“中级”“初级”“三十五岁以下”“三十六岁到四十五岁”“四十六岁以上”为自变量，进行最小二乘线性回归。

（3）因子分析

第一，对构成保费收入的各个变量提取公因子。

第二，对构成赔款支出的各个变量提取公因子。

（4）聚类分析

第一，使用构成保费收入的各个变量对各个财险公司进行聚类。

第二，使用构成赔款支出的各个变量对各个财险公司进行聚类。